UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE EDUCACIÓN
DEPARTAMENTO DE MÉTODOS DE INVESTIGACIÓN Y DIAGNOSTICO EN EDUCACIÓN

TESIS DOCTORAL

Propuesta metodológica para el estudio de la equivalencia entre dos versiones de una prueba: funcionamiento diferencial de versiones utilizando Propensity Score

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

Eva Jiménez García

DIRECTORES

José Luis Gaviria Soto
Rosaría Martínez Arias
Chantal Biencinto López

Madrid, 2017

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE EDUCACIÓN

DEPARTAMENTO DE MÉTODOS DE INVESTIGACIÓN Y DIAGNÓSTICO EN EDUCACIÓN

TESIS DOCTORAL

PROPUESTA METODOLÓGICA PARA EL ESTUDIO DE LA EQUIVALENCIA ENTRE DOS VERSIONES DE UNA PRUEBA: FUNCIONAMIENTO DIFERENCIAL DE VERSIONES UTILIZANDO PROPENSITY SCORE

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Eva Jiménez García

Dirigida por:

Dr. José Luis Gaviria Soto

Codirigida por:

Dra. Rosario Martínez Arias

Dra. Chantal Biencinto López

Madrid, 2016

PROPUESTA METODOLÓGICA PARA EL ESTUDIO DE LA EQUIVALENCIA ENTRE DOS VERSIONES DE UNA PRUEBA: FUNCIONAMIENTO DIFERENCIAL DE VERSIONES UTILIZANDO PROPENSITY SCORE

Tesis doctoral realizada por

Eva Jiménez García

Dirigida por:

Dr. José Luis Gaviria Soto

Codirigida por:

Dra. Rosario Martínez Arias

Dra. Chantal Biencinto López

UNIVERSIDAD COMPUTENSE DE MADRID
FACULTAD DE EDUCACIÓN

DEPARTAMENTO DE MÉTODOS DE INVESTIGACIÓN Y DIAGNÓSTICO EDUCATIVO

Madrid, 2016

A mi familia

AGRADECIMIENTOS

Me produce una enorme alegría escribir estas líneas, esto significa que llega el final de una etapa y el principio de otra. Además, tengo la suerte de poder agradecer, en estos párrafos, a todas esas personas tan importantes en mi vida y que han caminado a mi lado en esta autopista profesional, llena de paisajes preciosos y de algún que otro bache, que te hace ser más fuerte.

En primer lugar quiero dar las gracias a mis directores, Dr. José Luis Gaviria Soto y Dra. Rosario Martínez Arias. Gracias por compartir conmigo vuestra sabiduría, ha sido un verdadero placer empaparme de vuestras ideas y conocimientos. Gracias por vuestro entusiasmo y por dejarme aprender a vuestro lado, y hacerlo siempre con una sonrisa. Y a ti, mi directora y amiga, Dra. Chantal Biencinto, me has enseñado, guiado y ayudado. Me has hecho sonreír cuando apenas podía y me has dado fuerzas para conseguir esto.

Quisiera continuar agradeciendo a todos y cada uno de los profesores del departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE). Me habéis convertido en lo que soy, una docente e investigadora ilusionada por y con su trabajo. No puedo olvidarme de Miguel Serra, y su sonrisa, gracias por hacerlo todo tan fácil. Tote, mi compañera de despacho, gracias por todo el tiempo compartido y por estar siempre a mi lado ayudándome.

Por supuesto, gracias al grupo de investigación Medida y Evaluación de Sistemas Educativos (MESE), por dejarme formar parte de él. Ángeles Blanco, Covadonga Ruiz, Inmaculada Asensio, Xavier Ordóñez y por supuesto María Castro, habéis logrado inculcarme el valor que tiene la investigación, a vuestro lado solo es posible aprender y disfrutar del proceso. Gracias también a Elvira Carpintero y Mercedes García, por contar conmigo y dejarme participar en proyectos del Servicio de Orientación Universitario (SOU), por vuestro entusiasmo, vuestras ganas de innovar y vuestro excelente trabajo, que también han calado en mí.

Quiero dedicarle las siguientes líneas a mi compañera pero por supuesto amiga, Coral González. Gracias por todo, gracias por estar en los buenos y en los malos momentos, nunca olvidaré todo lo que has hecho por mí y todo lo que he aprendido a tu lado.

Enrique Navarro, Esther López y Eva Expósito (Equipo E), sois excelentes profesionales y amigos. Habéis estado siempre que os he necesitado, habéis sido mi mayor apoyo. Solo puedo decir de vosotros que, para mí, sois todo un ejemplo a seguir.

Gracias a Joaquín y Alicia, compañeros y excelentes personas que hicieron de mi estancia en el Instituto de Investigación y Desarrollo Educativo (Unidad de Evaluación Educativa, México) una aventura profesional y personal inmejorable. Nunca olvidaré la oportunidad que me brindaron de formarme y desarrollar competencias personales y profesionales imprescindibles para el desarrollo de mi labor docente e investigadora.

El rumbo de la vida va cambiando, y tuve que acudir a otra autopista profesional, la Universidad Europea de Madrid, donde actualmente ejerzo como docente. Gracias Bianca Thoilliez, Sarah Martín, Ana Moreno, Germán Gómez, Esther Moraleda, Rebeca Cordero, María Sánchez, Antonio Pinto, por darme esta magnífica oportunidad de cumplir mi sueño y sentirme tan valorada personal y profesionalmente a vuestro lado.

También quiero agradecer a todos y cada uno de mis amigos, tanto de Madrid como de Lagartera, y amigas de la carrera, a mis primos, tíos (en especial a Sagrario por la ayuda recibida), por mis ausencias, agobios y compromisos laborales. Solo sé que siempre habéis estado a mi lado, en los buenos y malos momentos, me habéis respetado y sobre todo comprendido, gracias por facilitarme el camino y por ayudarme a desconectar en momentos duros.

Por último, y no por ello menos importante, quiero agradecer a mi familia y a vosotros, M. ª Carmen y Luis Miguel, mis padres y mi querido hermano Alberto. Sois el motor de mi vida, sois los protagonistas de este resultado, porque gracias a vosotros soy lo que soy, personal y profesionalmente. Quiero agradeceros vuestra insistencia por formarme y por hacer las cosas bien, por darle valor a la perseverancia, constancia y esfuerzo en el trabajo. Todo ello me lo habéis transmitido con el ejemplo. No puedo olvidarme de nuestro lema, ese que nos trasmitiste, papá, y que perdurará en generaciones ulteriores, ese que nunca puede faltar en cada una de nuestras mesas de trabajo, escrito con nuestro puño y letra: “Querer es poder”. Con este trabajo espero y deseo que estéis tan orgullosos de mí como lo estoy yo de vosotros. Os quiero mamá, papá y hermano.

Y a ti, la pieza última del puzle, el comienzo de una nueva autopista juntos. Llegaste al final del proceso pero en el momento perfecto. Sin conocer en lo que estaba inmersa, me has respetado y has sabido comprender el ritmo frenético de esta profesión y cada una de mis ausencias. Me he encontrado con el mejor de los compañeros de viaje. Has sido capaz de ayudarme, escucharme y consolarme, me has transmitido tu optimismo y alegría. Gracias por hacerme reír como en la vida nadie lo había hecho, por hacerme feliz, por llenarme de amor y por caminar siempre juntos. Gracias por aparecer en mi vida. Te quiero, Berni: este éxito también es tuyo.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

En este trabajo se presenta un estudio comparativo entre dos versiones (papel y online) de una prueba de rendimiento en Comprensión Lectora y el efecto que tienen las dos versiones en la puntuación de cada ítem. Dicha prueba se enmarca dentro de la Evaluación de Diagnóstico de 2010-2011 en la Comunidad de Madrid, para estudiantes de 4º curso de Educación Primaria y 2º curso de Educación Secundaria Obligatoria.

Para el estudio comparativo se ha llevado a cabo un análisis descriptivo y psicométrico de cada ítem en ambas versiones, atendiendo a las medias, dispersiones, formas de distribución de las puntuaciones y fiabilidad. También se ha procedido a la validación atendiendo a la Teoría de Respuesta al Ítem y al estudio del supuesto de unidimensionalidad. Además, se han utilizado como medidas de comparación la prueba de igualdad de varianzas, la prueba T (Student), y la prueba Chi cuadrado (Pearson).

Igualmente se ha realizado una propuesta metodológica, "Funcionamiento Diferencial de Versiones" (DVF), basada en los estudios del Funcionamiento Diferencial de los Ítems y del Funcionamiento Diferencial de los Sujetos. Esta metodología se caracteriza por la utilización del *Puntaje de Propensión* o *Propensity Score* para conseguir un grupo de sujetos homogéneo y poder establecer comparaciones sobre la existencia o no del Funcionamiento Diferencial de Versiones. Para su estudio se han utilizado varios procedimientos, algunos de ellos de carácter descriptivo, como la Transformación del Índice de Dificultad (T.I.D.), Rajú, Lord, Estandarizado-Stand y Mantel–Haenszel, y otros, más potentes, como la Regresión Logística utilizando las puntuaciones TRI de los sujetos. Ante las comparaciones múltiples, para el estudio del DVF y para evitar los falsos positivos, se ajustaron los valores *p* mediante la corrección de Benjamin–Hochberg (1995).

Los resultados confirman la equivalencia entre ambas versiones (papel y online), tanto en Primaria como en Secundaria. Las distribuciones, los niveles de fiabilidad y la estructura factorial presentan valores semejantes. La media tiende a favorecer a la prueba en papel, pero el tamaño del efecto es mínimo, además de la no-existencia de Funcionamiento Diferencial de Versiones (existen ítems con DVF, pero la medida del efecto nos indica que el DVF es irrelevante en todos los ítems). Por todo lo expuesto, concluimos que ambas versiones son equivalentes.

ABSTRACT

This paper presents a comparative study of two versions (paper and online) of a Reading Comprehension Performance Test and the effect these two versions had on the score of each item. This test was included in the 2010-2011 Diagnostic Evaluation in the Community of Madrid for students in Year 4 of Primary Education and Year 2 of Compulsory Secondary Education.

To conduct the comparative study, a descriptive and psychometric analysis of each item in both versions was carried out, considering the means, dispersion, distribution of scores and reliability. Likewise, a validation was conducted following the Item Response Theory and the dimensionality assumption was studied. Furthermore, the comparison measures used were the test for equality of variances, Student's t-test as well as Pearson's chi-squared test.

Similarly, a "Differential Version Functioning" (DVF) method was conducted using Differential Item Functioning and Differential Subject Functioning. This method is characterised by the use of a *Propensity Score* to obtain a homogenous group of subjects and be able to establish comparisons on the existence, or not, of Differential Version Functioning. The study employed several procedures, some of them descriptive, including: Difficulty Index Transformation (DIT), Rajú, Lord, Standardised-Stand, Mantel–Haenszel, in addition to the more powerful Logistic Regression using the subjects' IRT scores. For the DVF study and to avoid false positives due to multiple comparisons, the *p* values were adjusted with the Benjamin–Hochberg procedure (1995).

The results confirm the equivalence between the two versions (paper and online) both in Primary and in Secondary Education. Distribution, reliability levels and factor structure were very similar. The mean tends to favour the paper test, but the effect size is minimal. In addition, there was no Differential Version Functioning (there were items with some DVF, but the size of the effect showed that DVF was irrelevant in all items). Therefore, we can conclude that both versions are equal.

INTRODUCCIÓN

Nunca consideres el estudio como una obligación, sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber.

Albert Einstein (1879-1955)

Cada vez es mayor la importancia que toman las evaluaciones de los sistemas educativos, y mucho más desde la llegada de pruebas internacionales a gran escala, de gran rigor, cuyos resultados, sustentados por el apoyo político, son indicadores del éxito o el fracaso de los diferentes sistemas educativos suponen indicadores de éxito (Leighton, 2012). Las evaluaciones a gran escala pueden definirse como instrumentos para el estudio de los sistemas educativos que nos permitirán llevar a cabo un diagnóstico del panorama y facilitarán, de ser necesario, la toma de decisiones para la mejora de la calidad educativa.

Los estudios internacionales *"ofrecen una información imprescindible sobre la situación de la educación y, por tanto, facilitan la toma de decisiones por parte de los responsables de las políticas educativas para mejorarlos"* (Roca, 2009, p.51). Estos estudios son diseñados con la intención de mejorar el conocimiento y comprensión de la situación educativa por parte de los diferentes países (Beller, 2013).

En palabras de Castro (2009) *"la evaluación de los sistemas educativos constituye una herramienta fundamental para conocer el progreso en educación* [...] *informando del estado actual del nivel académico de los estudiantes"* (p.1). Se recomienda la aplicación de estas pruebas en un gran número de países, con la intención de mejorar sus respectivos sistemas educativos, ya que permiten *"valorar la eficacia y el funcionamiento de las políticas educativas adoptadas, el análisis de la evolución del sistema y su comparación con otras realidades educativas"* (Maestro, 2006, p.315).

El objetivo de estas evaluaciones es *"proporcionar información analizada y valorada para tomar decisiones conducentes a la optimización del sistema evaluado; resumido en la siguiente proposición: investigamos lo que queremos conocer, evaluamos lo que queremos transformar"* (MESE, 2010, p.1).

El presente trabajo de investigación muestra los datos de la Evaluación de Diagnóstico Censal aplicada en la Comunidad de Madrid en 2010-2011. El marco normativo en el que se ampara la Evaluación de Diagnóstico es la Ley Orgánica 2/2006 de Educación, (LOE 2/2006, 3 de mayo), que establece en sus artículos 21 y 29 la obligatoriedad de la realización de evaluaciones de diagnóstico censales del sistema

educativo al finalizar el segundo ciclo de educación Primaria (4º Educación Primaria) y el segundo ciclo de Educación Secundaria (2º Educación Secundaria Obligatoria). Esta evaluación, competencia de las Administraciones Educativas, tendrá carácter formativo y orientador para los centros, e informativo para las familias y para el conjunto de la comunidad educativa.

En el ámbito de la Comunidad de Madrid, la realización de la Evaluación de Diagnóstico de cuarto curso de Educación Primaria y segundo curso de Educación Secundaria Obligatoria se prevé en el artículo 17 de la Orden 3319-01/2007 y la Orden 3320-01/2007 (20 de junio). Este artículo regula la implantación y organización de la Educación Primaria y Educación Secundaria Obligatoria, derivadas de la Ley Orgánica de Educación (LOE 2/2006, 3 de mayo), donde se establece que todos los estudiantes, al finalizar el segundo curso de Educación Primaria y el segundo curso de Educación Secundaria Obligatoria, habrán de realizar la prueba.

En la Evaluación de Diagnóstico las destrezas evaluadas son Matemáticas, Comprensión Lectora, y Lengua y Literatura, tanto para Primaria como para Secundaria (también se realiza un dictado en el caso de Primaria). En este trabajo se presentan los datos de la prueba de evaluación de dominios cognitivos basados en la Comprensión Lectora.

La peculiaridad de estas pruebas es el modo de aplicación, dado que se ofrece a los centros educativos la posibilidad de realizar la prueba en papel o informatizada. Esta posibilidad es consecuencia de los avances técnicos a los que hoy en día estamos expuestos.

Los test han cambiado y evolucionado, adaptándose a la sociedad tecnológica en la que nos vemos inmersos, con un gran número de programas informáticos y estadísticos que nos permiten realizar análisis de gran complejidad y solventar las nuevas necesidades sociales que la actualidad demanda (Abad, Olea, Ponsoda, y García, 2011; Olea, Abad y Barrada, 2010; Sireci y Zenisky, 2006).

Los ordenadores, cada vez más potentes, facilitan la "convivencia entre test y ordenadores", lo que abre nuevos caminos en el campo de la medición con test (Muñiz y Fernández-Hermida, 2010; Muñiz, 2012).

Vemos entonces que el ordenador ha pasado a tomar un papel de gran relevancia en el sistema educativo, trasladándose los test convencionales aplicados en papel y lápiz al soporte informático. En su gran mayoría, estos test son *"el resultado de virtualizar test convencionales en papel y lápiz con objeto de administrarlos, corregirlos y analizar los resultados mediante técnicas y programas informáticos"* (López-Mezquita, 2005, p.622).

Mención especial merece la aplicación de la tecnología en las evaluaciones internacionales a gran escala, como comenta Beller (2013) en su libro *"The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research"* (Capítulo 3: *"Technologies in Large-Scale Assessments: New Directions, Challenges, and Opportunities"*).

Cada vez son más las evaluaciones internacionales que trabajan con el modo de aplicación informatizado, como, por ejemplo:

International Computer and Information Literacy Study (ICILS, 2013): el primer estudio internacional que analiza el nivel de alfabetización en el uso del ordenador y el manejo de la información con diferentes fines (Fraillon, Schulz, Friedman, Ainley y Gebhardt, 2015).

Programme for International Assessment of Adult Skills (PIAAC, 2013): Programa Internacional para la Evaluación de las Competencias de la población adulta, cuyo cuestionario se aplica por ordenador y consiste en un test adaptativo informatizado.

Programme for International Student Assessment (PISA): no se mantiene al margen en lo que al modo informatizado de la pruebas se refiere (OCDE, 2013a). En la evaluación llevada a cabo hace unos meses, en 2015, contaron con una muestra de 40.000 estudiantes en España, y 525.000 en todo el mundo, que realizaron las

pruebas de Ciencias, Resolución de Problemas, Matemáticas y Comprensión Lectora por ordenador[1]. Los resultados todavía no están disponibles.

Hay que mencionar que PISA ya hacía años que se había sumado a la evaluación informatizada. Una prueba de ello es la *Electronic Reading Assessment* (PISA–ERA) (OCDE, 2009), centrada exclusivamente en la evaluación de la Lectura Digital (PISA 2012) (OCDE, 2013b). Ésta fue la primera ocasión en que se evaluó por ordenador la competencia en matemáticas, además de la comprensión lectora.

En este contexto vamos a estudiar una de las preocupaciones cruciales a la hora de informatizar un test convencional, determinar si el modo en que se aplican los test origina puntuaciones equivalentes, o si, por el contrario, es necesario cambiar el baremo de referencia dado que las diferencias entre uno y otro método afectan al rendimiento de los sujetos. De asumir esta equivalencia, podremos utilizar test informatizados en lugar de la versión en papel, y así beneficiarnos de las ventajas que conllevan dichas aplicaciones.

Para el estudio de la equivalencia entre dos versiones de una prueba es fundamental la existencia de grupos de sujetos equiparables que realicen la prueba en papel y online. Es en este punto donde surge la propuesta metodológica presentada en este trabajo: *Funcionamiento Diferencial de Versiones*, basada en el Funcionamiento Diferencial de Ítems y en el Funcionamiento Diferencial de Sujetos. Con motivo del sesgo de selección de la muestra objeto de estudio, se recurre a la metodología de Puntaje de Propensión (Propensity Score) para alcanzar grupos homogéneos y realizar adecuadamente el estudio de equivalencia.

El trabajo que aquí presentamos se agrupa en ocho capítulos, el primero de los cuales realiza un breve recorrido teórico sobre los test y sus modos de aplicación, así como sobre los beneficios y limitaciones que conlleva el uso de test informatizados frente a los test convencionales realizados en papel.

[1] Una simulación de los ítems por ordenador puede consultarse en los siguiente links: http://educalab.es/inee/evaluaciones-internacionales/preguntas-liberadas-pisa-piaac/pisa-por-ordenador. http://www.mecd.gob.es/inee/Preguntas-liberadas.html#PISA

El segundo capítulo recoge la necesidad de atender a las directrices para la adaptación de los test en función del modo de aplicación (convencional e informatizada): *"International Guidelines on Computer-Based and Internet Delivered Testing"*, desarrolladas por la International Test Commission (2005) y Standards for Educational and Psychological Testing (AERA, APA, y NCME, 2014). Estas directrices nos indican la importancia de realizar un estudio completo de las versiones, donde el estudio del Funcionamiento Diferencial de los Ítems es requisito primordial.

El tercer capítulo presenta la propuesta metodológica denominada *Funcionamiento Diferencial de Versiones (DVF)*, aplicando la metodología del Puntaje de Propensión (Propensity Score) para garantizar la equivalencia entre los grupos a comparar.

El capítulo cuarto, último de la fundamentación teórica, recoge los diferentes procedimientos que existen para la detección y el estudio del DVF, centrándose en el método de Regresión Logística utilizado en el presente trabajo.

Los siguientes capítulos corresponden al trabajo empírico llevado a cabo. En los capítulos quinto y sexto se presentan los problemas y las hipótesis, así como el diseño de la investigación realizada, lo que nos conduce al séptimo capítulo, en el que se recogen los resultados del estudio de la equivalencia entre las dos versiones (papel y online). En este capítulo podemos encontrar la demostración del uso de las Propensity Score para el emparejamiento efectivo de muestras, la validación atendiendo a la Teoría Clásica de los Test, la Teoría de Respuesta al Ítem, el estudio del supuesto de unidimensional, los contrastes de hipótesis para el estudio de la equivalencia y el posterior estudio del Funcionamiento Diferencial de Versiones.

Finalizamos con un capítulo destinado a la discusión y las conclusiones principales del trabajo. Por último, también se incluyen la bibliografía utilizada y los anexos.

Este trabajo y la Evaluación de Diagnóstico de 2010-2011 han sido posibles gracias a la Consejería de Educación, que asigna la competencia de la evaluación del sistema educativo de la Comunidad de Madrid a la Subdirección General de Evaluación, encargada de la elaboración, difusión, aplicación y corrección de las pruebas; y se ha contado a lo largo de todo el proceso de evaluación con el asesoramiento metodológico

del Grupo de Investigación *Medida y Evaluación de Sistemas Educativos*, de la Universidad Complutense de Madrid.

PARTE 1: FUNDAMENTACIÓN TEÓRICA

CAPÍTULO 1: La influencia de las innovaciones tecnológicas en los test

"Ser evaluado mediante un ordenador puede pronto llegar a ser incluso más natural que ser evaluado en papel".

(Davey, T. 2005, p.358)

1.1. Clasificación de los test

Con motivo de los avances técnicos actuales, los test han experimentado cambios y han evolucionado tanto en su formato como en su modo de aplicación, fruto de las tecnologías que, entre otras utilidades, impulsan la aplicación de los test por ordenador, la creación de bancos de ítems y el surgimiento de test adaptativos informatizados. A continuación presentamos una clasificación de los test surgida a raíz de los cambios propiciados por las innovaciones tecnológicas.

1.1.1. Test convencionales

Cuando hablamos de test convencionales nos referimos a los test tradicionales, en formato papel.

Son muchas las definiciones propuestas por diversos autores, entendiendo un test como "*el procedimiento sistemático para observar la conducta y describirla con la ayuda de escalas numéricas o categorías establecidas"* (Cronbach, 1990, p.32), o, en palabras de Renom (1992, p.581), "*un conjunto homogéneo y estandarizado de ítems cuyo objetivo es la evaluación cuantitativa bajo condiciones rigurosamente estandarizadas de sesgos y atributos psicológicos y educacionales"*. Martínez Arias entiende por test *"un reactivo que, aplicado a un sujeto, revela y da testimonio del tipo o grado de su aptitud, de su forma de ser o del grado de instrucción que posee"* (1995, p.31). Un test es un procedimiento para la evaluación y la obtención de una muestra sobre la conducta examinada en un dominio concreto, que es evaluado con anterioridad por procedimientos estandarizados (Standards for Educational and Psychological Test, APA, AERA y NCME, 1999) o "*procedimientos de recolección de información sobre un individuo o grupo y su construcción […] se basa en modelos psicométricos que permiten evaluar la calidad de la medida y dar garantías de la misma"* (Attorresi, et. al. 2009, p.179).

Tradicionalmente, estos test convencionales han sido los más utilizados, ya que suponían la única forma disponible de aplicar un test. Se caracterizan por la no informatización a lo largo de todo el proceso: la presentación de los ítems y la recogida de

datos se llevan a cabo a través de cuadernillos, es decir, se realizan en papel, al igual que sucede con la corrección de los test y el almacenamiento de datos, que se producen manualmente.

Los test han sido utilizados durante siglos; si nos remontamos a sus orígenes, en China, el emperador estableció un Servicio Civil de Evaluación con la finalidad de examinar la valía de sus oficiales en el trabajo que desarrollaban. Se instauró un programa que evaluaba las competencias profesionales de los oficiales del gobierno, es decir, las capacidades de estos en áreas como la agricultura, la contabilidad, la lectura y la escritura.

Hasta el siglo XVI no se comenzaron a usar test para la evaluación de estudiantes, siendo los jesuitas los primeros que evaluaron a sus alumnos con este método (Martínez-Arias, 1995). Es precisamente en este siglo cuando, producto del desarrollo, se da una mayor demanda educativa, lo que ocasiona un aumento de la necesidad de comprobación de las valías individuales y, por consiguiente, un auge de las normas sobre la utilización de exámenes escritos (Gil, 1992).

El origen de los actuales test deriva de los intentos de medir las diferencias individuales de los sujetos, y van unidos a nombres tales como:

Galton, cuya mayor contribución fue la utilización de la metodología estadística en el análisis de datos para la detección de las diferencias individuales, a través de herramientas como la correlación, métodos de escalamiento psicológicos y medidas estándar.

Catell, discípulo de Galton, fue el primero en utilizar el término *test mental* en el artículo *"Mental Test and Measurements"* en 1890, cuya finalidad era medir la inteligencia.

Otra figura importante en la historia de los test es Stern, quien dio origen a un nuevo concepto, el *Cociente Intelectual (CI),* que cuantificaba la inteligencia por medio de un indicador.

Mención especial merece la obra de Spearman *"General Intelligence Objectively Determined and Measured"*, de 1904, que fue determinante en el campo educativo, ya que en ella trató de explicar las deficiencias de los métodos y teorías empleados para la medición de la inteligencia, indicando que no solo los test mentales deben correlacionarse con la inteligencia, sino que hay que ampliarlos a otros ámbitos como el educativo y académico.

Un número bastante amplio de pruebas fueron desarrolladas durante las dos guerras mundiales; en concreto, *"poco antes de la Primera Guerra Mundial se da un reconocimiento institucional al papel de los test en el procedimiento diagnóstico"* (Thompson y Sharp, 1988, citado por Navas, 1999, p.2), aunque, realmente, no es hasta 1920 cuando surge la primera batería estandarizada para medir el rendimiento y la evaluación de distintos ámbitos educativos.

Hoy en día se trata de una práctica bastante habitual, siendo tradicionalmente la manera más común de recabar información sobre los estudiantes en centros educativos (rendimiento, actitudes, perfiles profesionales), al igual que en la enseñanza superior (pruebas de acceso) o en pruebas psicológicas y psicopedagógicas. Actualmente, con nuevos adelantos tecnológicos, se ha potenciado la utilización de test informatizados, que facilitan y agilizan el proceso de evaluación.

1.1.2. Test informatizados

Cuando se habla de test informatizados se hace referencia a todas las pruebas cuyo formato de administración es por medio del ordenador. Las evaluaciones informatizadas no hubieran sido posibles sin los *"desarrollos e innovaciones que se han producido en la medición educativa, sobre todo en los modelos teóricos de la Teoría de Respuesta al Ítem"* (Belloch, 2011, p.1), fundamentalmente porque se necesita un soporte informático que requiere programas estadísticos para la estimación de parámetros y ajuste de los modelos TRI. Cada vez es mayor el número de programas informáticos y estadísticos que nos permiten realizar investigaciones y análisis de gran complejidad (Rodríguez y Martínez, 2003).

Existen diferentes tipos de test informatizados, siendo los más comunes los convencionales, que engloban una informatización sin requerimiento de algoritmos para la selección de ítems. Pero también existen los test adaptativos informatizados, en los que sí se producen algoritmos para la selección. Para entender estos tipos de test es necesario ser conscientes del cambio tan relevante que han ido sufriendo a lo largo del proceso, fruto de las innovaciones tecnológicas que actualmente nos rodean; por esta razón presentamos en las siguientes líneas el proceso de informatización de los test.

En un principio, los ordenadores eran utilizados para corregir los test convencionales realizados en papel, así como para elaborar de forma objetiva informes sobre los resultados obtenidos, como la corrección automática del test Minnesota Multiphasic Personality Inventory – MMPI (Inventario Multifásico de Personalidad de Minnesota) y la elaboración automática de informes (Bartram, 2006).

En los años cincuenta se produjo un avance muy importante en lo relativo a la introducción de datos, pues las tarjetas fueron sustituidas por la aplicación de hojas de respuesta electrónica y hojas de lectura óptica. Fue en esta década, en Estados Unidos, cuando empezaron a utilizarse los ordenadores con fines educativos (Chapelle, 2001).

En los años setenta, Lord comenzó a utilizar ordenadores para el diseño y la aplicación de test en grupos de estudiantes; lo que en los setenta era algo novedoso hoy se considera natural. Durante los años siguientes los ordenadores comenzaron a tener presencia en la educación, y a partir de ese momento su uso educativo aumentó vertiginosamente. Ante este panorama, las escuelas aumentaron el número de ordenadores por estudiante, pasando de un ordenador por cada 125 estudiantes en 1983 a un ordenador por cada nueve en 1995 (Russell, 1999).

En Estados Unidos se produjo un desarrollo de sistemas automatizados de interpretación de test convencionales, que producen potentes programas informáticos capaces de recoger bases de datos de gran tamaño (Molina, Sanmartín y Pareja, 2000; Seisdedos, 1999).

Todavía en los setenta, y como resultado de la potencia cada vez mayor que van desarrollando los ordenadores, se realiza por primera vez la aplicación informatizada de test de aptitudes e inteligencia tales como WAIS (Wechsler, Coalson y Raiford, 2008), el Test de las Matrices Progresivas de Raven (Casé, Neer, Lopetegui, Doná, Biganzoli y Garzantini, 2015), el Test de Figuras Enmascaradas (Witkin, Oltman, Raskin y Karp, 1987) y el Slosson Intelligence Test (Slosson, Nicholson y Hibpshman, 1991).

En la década siguiente se produjo un desarrollo importante de los test informatizados, lo que llevó a la informatización de muchos de los test convencionales. En el ámbito educativo encontramos los test de Habilidades y Conocimientos Básicos informatizados (Backhoff, Ibarra y Rosas, 1994); test informatizados para la certificación de títulos y para las pruebas de nivel elaboradas por el Educational Testing Service. Atendiendo a los niveles educativos objeto de estudio en este trabajo, destacamos también la Batería Escolar Informatizada de TEA para la evaluación de escolares de 4º de Educación Primaria: EFAI-1 (Aptitudes), FI-R (Evaluación de las aptitudes perceptivas y de atención), ECL-2 (Evaluación de la comprensión lectora), así como para 2º de Educación Secundaria: EFAI-3 (Aptitudes), C. L. (Evaluación de aptitudes perceptivas y de atención), BFQ-NA (Cuestionario de personalidad), e IPP-R (Intereses y preferencias profesionales).

1.1.2.1. Test convencionales informatizados (1ª Generación)

Los avances en el terreno educativo y evaluativo suponen un salto vertiginoso que propicia el desarrollo de nuevas formas de aplicación de los test, en los que el ordenador toma un papel primordial (Drasgow, Luecht y Bennet, 2006).

Como señala Davey (2005, p.358), *"ser evaluado mediante un ordenador puede pronto llegar a ser incluso más natural que ser evaluado en papel"*, y cada vez es menor el uso de test soporte papel y mayor la utilización del "ratón y monitor" o "test informatizados" (Molina, Sanmartín y Pareja, 2000). Por esta razón se comenzaron a trasladar los test convencionales aplicados en papel y lápiz al soporte informático, convirtiéndose así en la *"primera generación"* de evaluaciones asistidas por ordenador

(Bunderson, Inouye y Olsen, 1988), dando lugar a los test convencionales informatizados, también conocidos como la traducción a soporte informático de test aplicados en papel.

Estos test son *"la aplicación primera y más natural de la informática al ámbito de los test convencionales de papel y lápiz en un ordenador y aplicados a través del teclado y la pantalla"* (Muñiz y Hambleton, 1999, p.24).

Existen diferentes niveles de informatización de un test; a continuación se recogen un conjunto de tareas cuyos procesos de informatización han evolucionado con el tiempo (tabla 1.1). Una primera fase es una informatización parcial, en la que los ítems son aplicados en papel y lápiz y corregidos con ordenador, para lo cual es necesario realizar la introducción de los datos, bien manualmente o con lectoras ópticas. La segunda fase consiste en la informatización completa del proceso: los ítems y sus instrucciones son presentados y respondidos por ordenador.

Tabla 1.1.
Tareas en el proceso de informatización de un test

Tarea	**No informatizado**	**Informatizado**	
		1ª fase	**2ª fase**
Presentación del test	Cuadernillo (en papel)	Cuadernillo (en papel)	En pantalla del ordenador
Recogida de datos	En cuadernillo	Mediante hoja de respuestas	En el ordenador, por medio del teclado y/ o ratón
Corrección del test	Manualmente, con plantillas	Lectoras ópticas	Con programación
Almacenamiento de datos	Archivadores	Ficheros de texto	Bases de datos y/o hojas de cálculo
Análisis de datos	Cálculos manualmente	Programación	Programación
Informe de resultados	Manualmente	Informe a ordenador	Informe a ordenador

Fuente: Adaptación Seisdedos (1999)

Por todo ello, un test puede ser considerado informatizado en sentido estricto si el ordenador es utilizado como medio para presentar los ítems, responderlos, analizarlos y para interpretar los resultados obtenidos (Olea y Hontangas, 1999).

Un apartado especial merecen los test informatizados cuya aplicación se lleva a cabo a través de internet, es decir, los test online, puesto que son los instrumentos utilizados en este estudio. Son test en los que se requiere de una conexión a internet y de un servidor para su presentación y almacenamiento de datos. Este tipo de test presenta algunas ventajas, como puedan ser el abaratamiento de costes, el acceso directo a bases de datos o el acceso selectivo a la información deseada (Olea, Abad y Barrada, 2010).

Para llevar a cabo la aplicación online, que requiere un acceso a internet, es imprescindible garantizar la seguridad y la protección de los datos (Bartram y Hambleton, 2006). Ante esta necesidad se ha requerido la creación de unas directrices para el diseño y la elaboración correcta de los test informatizados, habiendo un apartado especial en International Test Commission (2005) para las aplicaciones vía internet, recogido en el capítulo siguiente de del trabajo (Capítulo 2).

1.1.2.2. Test Adaptativos Informatizados – TAIs (2ª Generación)

Los Test Adaptativos Informatizados[2] (de ahora en adelante TAIs) son la *"segunda generación"* de las evaluaciones asistidas por ordenador (Bunderson, Inouye y Olsen, 1989). Se trata de *"una prueba, construida para fines de evaluación psicológica o educativa, cuyos ítems se presentan y responden mediante un ordenador, siendo su característica fundamental que se va adaptando al nivel de competencia progresivo que va manifestando la persona"* (Olea y Ponsoda, 2013, p.5).

El procedimiento de los TAIs se basa en la selección de los ítems uno a uno en función del nivel de habilidad que tenga el sujeto, para lo cual se utiliza un algoritmo en

[2] Acrónimo en inglés de la expresión "Computerized Adaptive Test" (CAT)

el que se seleccionan ítems más fáciles o más difíciles en función del acierto o error del ítem, obteniendo un mejor ajuste y una mayor precisión.

Estos test están tomando cada vez mayor relevancia; un ejemplo de ello son los test adaptativos utilizados para la admisión en centros educativos como el Law School Admission Test (LSAT), la prueba matemática para adultos (MATHCAT) o la prueba para evaluar la comprensión de textos (COMTEX). También son dignos de mención los test adaptativos informatizados para realizar pruebas de nivel de idiomas: el Test of English as a Foreign Language (TOEFL), el Graduate Management Admissions Test (GMAT), el Graduate Record Exam (GRE) o el Test informatizado para la evaluación de razonamiento secuencial y la inducción (TRASI).

La diferencia entre los test convencionales informatizados y los test adaptativos informatizados radica en que los primeros no requieren de algoritmos de selección de los ítems, ya que se presentan a todos los sujetos exactamente los mismos ítems; por otro lado, tampoco requieren análisis psicométricos complejos (López-Mezquita, 2005). Todo lo contrario sucede con los TAIs, que requieren algoritmos complejos, dado que no se conoce *a priori* el orden que van a tener los ítems.

Resumiendo, estos test se caracterizan por contar con bancos de ítems cuyas propiedades son conocidas y donde se establece el procedimiento adecuado para seleccionar y lograr la adecuación de los ítems en función de la habilidad de cada sujeto (Schade, Hernández, y Elgueta, 2005). Son, por tanto, fruto de la existencia de los bancos de ítems, de donde se seleccionan los más apropiados en función de la respuesta del examinado (Molina, Sanmartin y Pareja, 1999).

En el caso que nos ocupa, nuestro banco de ítems, como lo denominan Renon y Doval (1999), consistirá en el diseño de una matriz de especificaciones que incluya conocimiento y contenido, así como la representación de ítems que combinen la selección de contenidos y de objetivos óptimos.

1.1.2.3. Evaluación Continua (3ª Generación)

Para Bunderson, Inouye y Olsen (1989) la *"tercera generación"* es la evaluación continua o "continuous measurement". Esta evaluación, como su propio nombre indica, es continua y versa sobre la estimación y el pronóstico a través de algoritmos matemáticos, basados en los cambios habidos a lo largo de la trayectoria del aprendizaje curricular del estudiante.

Lo que distingue esta evaluación de otras es que asume la posibilidad de calcular la trayectoria del aprendizaje de los evaluados de forma significativa, ya que trata de pronosticar el aprendizaje partiendo de un nivel concreto de capacidad del evaluado hasta otro nivel específico que alcanzará en un futuro (López-Mezquita, 2005).

1.1.2.4. Evaluación Inteligente (4ª Generación)

La *"cuarta generación"* para Bunderson, Inouye y Olsen (1989) es la evaluación inteligente o "intelligent measurement". Esta evaluación, a través de técnicas de inteligencia artificial, *"produce, interpreta y genera perfiles de los resultados del estudiante con base en conocimientos y procedimientos de inferencia"* (Backhoff, Ramírez y Dibut, 2005, p.20)

Por tanto, esta evaluación trata de sugerir los estilos de aprendizaje y los contenidos acordes a la etapa estimada en la que se encuentra el alumno (López-Mezquita, 2005).

Todavía no se están implantando en el campo educativo ni la tercera ni la cuarta generación de test informatizados, debido a la gran cantidad de variables que influyen en el aprendizaje.

1.2. Beneficios y limitaciones de la informatización de los test

La utilización del ordenador para llevar a cabo evaluaciones supone unos beneficios frente a los test aplicados en formato papel y lápiz. Un ejemplo de ello es el reciente estudio llevado a cabo por Arachi, Dias y Madanayake (2014), en el que el 85% de los estudiantes preferían las pruebas realizadas por ordenador, ya que este formato les resultaba más cómodo, fácil y accesible. Si atendemos a lo que Fulcher (2000) nos sugiere, los ordenadores nos facilitan diferentes tareas, desde el proceso de elaboración del test hasta el informe final.

Son muchas las ventajas que nos ofrecen los ordenadores frente a las pruebas llevadas a cabo en papel (Drasgow y Olson-Buchanan, 1999); a continuación resumimos algunas de ellas.

Diseño de los test

Gracias a los ordenadores es posible crear test con diseños gráficos que faciliten la atención y potencien la motivación de los evaluados.

Los ítems han sufrido cambios en su edición con la llegada de las innovaciones tecnológicas, ya que éstas han abierto canales visuales y auditivos importantes, facilitando la medición y la posibilidad de aplicar estas pruebas a personas con patologías. Esto ocasiona que con los test informatizados sea posible medir determinados procesos imposibles de llevar a cabo con los test en papel y lápiz (Bartram, 2006).

Muñiz y Hambleton (1999, p.24) indican que *"el ordenador va a permitir utilizar ítems más complejos y cercanos a la realidad, que incluso se puede simular, lo que conllevaría posibles mejoras en la validez predictiva, al acercarse más el test a la realidad criterial que se pretende medir"*.

Construcción de test

Otra ventaja que implica la utilización de la informatización en los test es la posibilidad de diseñar los test de manera automatizada. Actualmente, para la elaboración de los test se hacen cada vez más necesarios programas informáticos que ensamblen los ítems con las propiedades psicométricas que se determinen. Para ello, Garcés, Sepúlveda y Riquelme (2014) mencionan algunos programas con óptimas características para la consecución de este objetivo, como puedan ser Quiz Press, iTest, Hot Potatoes, Wondershare QuizCreator, TestGIO Porfesor, TestGen 7.0, o ExamView. En Internet hay una gran variedad de programas y plantillas que permiten diseñar los test de forma rápida y fácil.

El diseño de test utilizando las innovaciones tecnológicas permite elaborar test combinando aspectos visuales y auditivos, lo que resulta en réplicas virtuales adaptadas a contextos reales (Van den Branden, Depauw y Gysen, 2002).

Aplicación del test

Durante su aplicación, los test informatizados permiten establecer controles que los procedimientos convencionales no permitían (Belloch, 2011; Molina, Sanmartín y Pareja, 2000). Por ejemplo, es posible fijar el tiempo de exposición de las preguntas, con lo que se consigue un control considerable, garantizando las mismas condiciones para todos.

Existe menor probabilidad de copia, en concreto cuando se utilizan test adaptativos informatizados, en los que, como su nombre indica, la presentación de los ítems se realiza de forma aleatorizada. Estos test suelen presentar un ítem por pantalla, lo que hace que sea más ameno que en papel, al presentarse a menudo este último formato en varias hojas.

Todo lo anterior supone un menor tiempo de aplicación, lo que se traduce en un menor número de ítems y, por ende, menores costes y mejores estimaciones.

Almacenamiento de datos

Gracias al soporte informático evitamos errores relacionados con la introducción de datos, tanto por parte del investigador (al introducir los datos en el ordenador) como por parte del sujeto (al responder en la hoja de respuestas).

Por tanto, el ordenador es un instrumento de gran potencial para el almacenamiento de datos, los cuales quedan automáticamente introducidos, facilitando esta tarea y otras relacionadas con la obtención de los resultados, análisis de datos y realización de informes.

Además de almacenar las respuestas a los ítems, los ordenadores permiten obtener otros registros, recabando información que con los test tradicionales es inviable obtener. Un ejemplo de ello son los registros del tiempo empleado en responder a cada uno de los ítems, las pulsaciones en el teclado o el número de respuestas correctas e incorrectas, todo ello guardado de manera automática (Alderson, 2000).

Elaboración de informes automáticamente

Como resultado de lo anterior, es decir, de la obtención de la puntuación de forma automática, es lícito que se pueda automatizar la elaboración de los informes y su debida interpretación de los datos (Muñiz y Hambleton, 1999; Schade, Hernández y Elgueta, 2005).

Estos informes son elaborados rápidamente, sin necesidad de invertir tiempo introduciendo los datos en un programa estadístico para posteriormente llevar a cabo los análisis necesarios para elaborar un informe, como sucede con los test en papel y lápiz, ya que con los test informatizados este proceso se realiza automáticamente.

Aún con todas estas facilidades, los ordenadores nunca deben sustituir la tarea diagnóstica del pedagogo o psicólogo, sino que deben servir de ayuda y como comienzo para desarrollar el diagnóstico oportuno y una elaboración correcta del informe, siguiendo la ética y el juicio profesionales.

A pesar de todas las ventajas presentadas en párrafos anteriores, existen algunas limitaciones a considerar, algunas de las cuales hacen referencia a aspectos personales que pueden afectar los resultados.

Principalmente se refieren a la ansiedad que puede ocasionar para el evaluado la realización de la prueba en ordenador, por la escasa familiaridad de éste con el medio (Goldberg y Pedulla, 2002; Pomplun, Ritchie y Custer, 2006), característica también compartida con los TAIs, ya que no ofrecen la posibilidad de corregir la respuesta una vez contestada.

A diferencia de los TI, los TAIs seleccionan los ítems atendiendo al nivel de rango del evaluado, lo que hace que éste no tengan que enfrentarse a ítems tan difíciles como para que le lleven a abandonar la prueba fruto de la frustración; o, por el contrario, tan fáciles que le aburran.

Otros inconvenientes son los relativos a las características propias de los ordenadores, concretamente a aspectos como el tamaño de las pantallas (que no siempre es el mismo), las condiciones (que no siempre son las deseadas, como la resolución o el tamaño de la fuente), problemas con el software de navegación y configuración, etc. Todo ello puede llegar a influir en los resultados (McKee y Levinson, 1990; Noyes y Garland, 2008).

En el caso concreto de los TAIs, el inconveniente principal es el gran tamaño del banco de ítems necesario para ejecutar dicha aplicación, ya que implica mucho tiempo y trabajo, además de resultar bastante costoso.

CAPÍTULO 2: Directrices para la adaptación de test informatizados y estudio de la equivalencia

"Todo lo que se hace se puede medir, sólo si se mide se puede controlar, sólo si se controla se puede dirigir y sólo si se dirige se puede mejorar".

Mendoza, P. (2001)

2.1. Introducción

Son muchos los avances que se han dado en el campo de los test y la evaluación con la llegada de las nuevas tecnologías, que han hecho del ordenador parte integrante del proceso evaluativo.

Considerando la relevancia que están tomando estos test, es fundamental garantizar que la adaptación de los test convencionales a los test informatizados sea la correcta. Para ello, existen unas directrices que nos van a guiar en el proceso y que presentamos en las siguientes líneas.

2.2. Directrices para adaptación de test informatizados

Dada la relevancia que las aplicaciones informatizadas están teniendo en la actualidad, la American Psychological Association Committee on Professional Standards y el Committee on Psychological Test and Assessment (1986), en su trabajo titulado *"Guidelines for computer-based test and interpretations"* presentan algunas normas para el uso correcto de los test informatizados.

Abordaremos trabajos más recientes como los *"Standards for Educational and Psychological Testing"* (AERA, APA, y NCME, 2014), así como las directrices de carácter internacional dirigidas especialmente a las pruebas aplicadas por ordenador/internet: "International Guidelines on Computer-Based and Internet Delivered Testing", de la International Test Commission (2005).

Estas directrices marcan las pautas a seguir para elaborar pruebas de forma adecuada. Las normas tienen carácter internacional, ya que todo el colectivo aúna fuerzas para lograr alcanzar buenas prácticas en lo que a adaptación y uso de los test se refiere.

Su finalidad se resume en las siguientes ideas (ITC, 2005):

"to produce a set of internationally developed and recognised guidelines that highlight good practice issues in computer-based (CBT) and Internet-delivered testing" (p.2).

"to raise awareness among all stakeholders in the testing process of what constitutes good practice" (p.2).

"These guidelines are intended to complement the ITC Guidelines on Test Use (2001), with a specific focus on CBT/Internet testing" (p.2).

La finalidad principal de las mencionadas directrices radica en producir, adaptar y crear conciencia de la necesidad de unas pautas reconocidas a nivel internacional, que manifiesten las buenas prácticas cuando se lleven a cabo pruebas por ordenador/internet.

Para llevar a cabo el proceso de elaboración de estas directrices, como así se recogen en ITC (2005), se utilizaron varios procedimientos que permitieron la recogida de información.

El primero de ellos es el referido a la búsqueda de bibliografía relevante sobre el tema, como pueda ser la Association of Test Publishers–ATP (2002), Bartram (2001, 2002), el British Standards' Institute–BSI (2001), o el British Psychological Society Psychological Testing Centre (2002), entre otros.

Por otro lado, se realizó una encuesta a los editores de pruebas en el Reino Unido y, por último, una Conferencia que permitió el intercambio de ideas entre expertos en el campo de los test informatizados.

Todo este material, y su pertinente revisión, permitieron crear una base consistente para la elaboración de las directrices sobre los test informatizados y por internet.

Las conclusiones fundamentales a las que llegaron se enmarcan dentro de cuatro grandes temas: Tecnología, Calidad, Control y Seguridad.

1. Tecnología (atiende a los aspectos técnicos como el hardware y software):

> *"Technology – ensuring that the technical aspects of CBT/Internet testing are considered, especially in relation to the hardware and software required to run the testing"* (ITC, 2005, p.3).

Tanto el hardware como el software, para su correcta utilización, requieren una descripción sobre el navegador, así como pruebas pertinentes para comprobar la plataforma donde se desarrollará el test (AERA, APA y NCME, 2014).

En relación a este punto, adjuntamos en el anexo 1 la guía para la aplicación online, redactada por el equipo MESE, que garantiza la correcta aplicación de la prueba.

AERA, APA y NCME (2014), en el capítulo 4, *"Developing Procedures and Materials for Administration and Scoring"* (p.83), mencionan las mismas ideas. Los procedimientos de administración deben ser consistentes, donde se especifiquen todo tipo de requisitos del hardware (velocidad del procesador, teclado, ratón, resolución de pantalla, conexión a internet) y software (bloqueo del acceso).

2. Calidad (del material y la correcta puesta en práctica de la prueba):

> *"Quality – ensuring and assuring the quality of testing and test materials and ensuring good practice throughout the testing process"* (ITC, 2005, p.4).

En relación a esta directriz, para la presentación del material hay que asegurarse de que la pantalla tenga la resolución adecuada y que disponga exclusivamente de la información necesaria para la prueba, que no esté sobrecargada. En lo relativo al formato del material, cabe mencionar a Kyllonen (1991, 1994 y 1996) cuando se centra en la estandarización del uso de colores (por ejemplo, para el título, blanco sobre negro; para las instrucciones, blanco sobre cyan; para los ítems, amarillo sobre azul, entre otros).

En lo que respecta a las instrucciones para la realización del test, de nuevo citamos a Kyllonen (1991, 1994 y 1996), pues los test informatizados y los convencionales tienen diferencias que han de mostrarse en las instrucciones. Las instrucciones deben reelaborarse y adaptarse a las condiciones de la administración informatizada, presentarlas en pantalla con un ejemplo del tipo de preguntas, el procedimiento para pasar a la siguiente pregunta, revisión las respuestas y cómo finalizar la prueba. En contraposición, en los test convencionales las instrucciones se presentan en papel y en una sola hoja.

AERA, APA, y NCME (2014) de nuevo recoge estas ideas en su capítulo 4, *"Developing Procedures and Materials for Administration and Scoring"* (p.83), donde alude a la importancia de informar a los sujetos de cómo se va a realizar la prueba (por ejemplo, si pueden o no volver hacia atrás para ver los ítems contestados, cómo navegar por la prueba, etc.).

3. Control (autenticidad del evaluado):

> *"Control – controlling the delivery of tests, test-taker authentication and prior practice"* (ITC, 2005, p.4).

4. Seguridad: este aspecto se desarrolla en profundidad en el trabajo llevado a cabo por ITC (2005), *"Guidelines on the Security of Test, Examinations, and Other Assessments"*, en el que se pretende garantizar la protección de los datos, la confidencialidad y el protocolo de seguridad.

> *"Security – security of the testing materials, privacy, data protection and confidentiality·* (ITC, 2005, p.4).

Estas directrices (control y seguridad) están interrelacionadas, dado que para garantizar la seguridad y la protección de datos es necesario llevar a cabo un control de la aplicación.

Cuando se utilizan test informatizados online es necesario que la aplicación sea supervisada para garantizar la seguridad, la protección de datos y evitar la suplantación de identidad; para ello es conveniente la asignación de contraseñas de acceso que controlen el cumplimiento de las condiciones necesarias para llevar a cabo la aplicación (Bartram y Hambleton, 2006; Becker y Bergstrom, 2013; Csapó, Ainley, Bennett, Latour y Law, 2012; ITC, 2005; Muñiz, Elosua y Hambleton, 2013). Por esto, en nuestro estudio, a cada centro y a cada alumno les fue asignada una clave para poder realizar el test y garantizar la protección de datos.

Las pruebas pueden ser administradas de diversas formas, según el nivel de control que los investigadores acometan. Bartram (2001) y el ITC (2005) proponen:

- Modo abierto (open mode): en esta aplicación no hay ningún tipo de supervisión ni control, es decir, el acceso es abierto, sin restricciones, por lo que no hay garantía de autenticidad de la identidad del evaluado.

- Modo controlado (controlled mode): similar al anterior, sin supervisión, pero para la realización de la prueba se requieren un nombre de usuario y una clave de acceso.

- Modo de supervisión (supervised mode): en este modo existe supervisión y control sobre posibles problemas; se garantiza un seguimiento por parte del supervisor de que la prueba se completó correctamente, así como la autenticidad de la identidad de los usuarios.

- Modo logro (managed mode): este es un modo en el que hay un alto nivel de supervisión, ya que se asume el control sobre el entorno, llevando a cabo pruebas sobre este tipo de test.

En el caso que nos concierne estamos ante el modo de supervisión, donde existe un supervisor (un profesor del centro) que lleva a cabo un seguimiento de la prueba y garantiza la seguridad y la protección de datos mediante claves.

Además, es necesario establecer mecanismos para la prevención y detección de errores. Para ello, durante la ejecución del test se impidió el funcionamiento de teclas innecesarias para cumplimentar el test, así como el bloqueo del acceso a internet, desde el comienzo de la prueba hasta su completa finalización.

Tras multitud de revisiones y versiones[3], se elaboraron finalmente las directrices actuales (ITC, 2005). Concretamente, éstas giran en torno a cuatro aspectos, recogidos en la tabla 2.1., con una presentación más detallada de las directrices concretas de cada uno de estos sectores (tecnología, calidad, control y seguridad) en los anexos 2, 3, 4 y 5.

Tabla 2.1
Directrices relacionadas con los aspectos tecnológicos, de calidad, control y seguridad

TECHNOLOGY	*Give due regard to technological issues in Computer-based (CBT) and Internet Testing* a. Give consideration to hardware and software requirements b. Take account of the robustness of the CBT/Internet test c. Consider human factors issues in the presentation of material via computer or the Internet d. Consider reasonable adjustments to the technical features of the test for candidates with disabilities e. Provide help, information, and practice items within the CBT/Internet test
QUALITY	*Attend to quality issues in CBT and Internet testing* a. Ensure knowledge, competence and appropriate use of CBT/Internet testing. b. Consider the psychometric qualities of the CBT/Internet test. c. Where the CBT/Internet test has been developed from a paper and pencil version, ensure that there is evidence of equivalence. d. Score and analyse CBT/Internet testing results accurately. e. Interpret results appropriately and provide appropriate feedback.

[3] Muñiz y Hambleton (1999), llevaron a cabo una traducción al español de las directrices para la correcta utilización de los test informatizados que puede verse en el anexo 6 de la tesis.

Tabla 2.1
Directrices relacionadas con los aspectos tecnológicos, de calidad, control y seguridad (continuación)

CONTROL	*Provide appropriate levels of control over CBT and Internet testing* a. Detail the level of control over the test conditions. b. Detail the appropriate control over the supervision of the testing. c. Give due consideration to controlling prior practice and item exposure. d. Give consideration to control over test-taker's authenticity and cheating.
SECURITY	*Make appropriate provision for security and safeguarding privacy in CBT and Internet testing* a. Take account of the security of test materials b. Consider the security of test-taker's data transferred over the Internet c. Maintain the confidentiality of test-taker results

Fuente: ITC (2005)

Es necesario cumplir con cada una de las directrices marcadas, para garantizar la igualdad entre versiones en el estudio de la equivalencia.

2.3. Estudio de equivalencia

Recurriendo a AERA, APA, y NCME (2014) y a sus estándares, en el capítulo 9: "*The rights and responsibilities of test users"*, se hace mención a los efectos significativos sobre la validez de los resultados, que pueden ocasionar los cambios en el modo de aplicación de las pruebas. El constructo evaluado puede verse modificado y por este motivo, es imprescindible realizar un estudio de la validez en ambas versiones, para garantizar su equivalencia (estándar 9.9.).

Standard 9.9:

"When a test user contemplates an alteration in test format, mode of administration, instructions, or the language used in administering a test, the user should have a sound rationale and empirical evidence, when possible, for concluding that the reliability/precision of scores and the validity of

interpretations based on the scores will not be compromised" (AERA, APA, y NCME, 2014, p.144).

Son muchos los autores que han trabajado al respecto, sugiriendo que el punto de partida para el estudio de la equivalencia de los test, sea la necesidad del cumplimiento de las mismas características; es decir, es indispensable que los test informatizados dispongan de las mismas condiciones que los test convencionales (Csapó, Ainley, Bennett, Latour y Law, 2012[4]).

Muñiz y Hambleton (1999), señalan que para garantizar la equivalencia, es necesario que los ítems en ambas versiones tengan un orden similar y que presenten semejantes medias, dispersiones y formas de distribución de las puntuaciones. En la misma línea, Bergstrom y Lunz (1999) indican que una de las características evidentes cuando hablamos de equivalencia es, la adecuada distribución de los parámetros de la dificultad de los ítems a lo largo del rango medido.

Numerosos autores, centran su atención en la necesidad de que el constructo medido y la estructura factorial, sea la misma en ambos grupos (Barbero, Vila, Holgado, 2008; Hambleton, 2005 y Van de Vijer y Leung, 2000).

Retomando los estándares, en el capítulo 1: *"Validity"*, se hace referencia a la validez de las pruebas y en concreto a la validez de la estructura interna, donde se especifica la importancia del estudio del funcionamiento diferencial de los ítems (AERA, APA, y NCME, 2014).

Atendiendo a los estándares mencionados, la metodología empleada para el análisis del sesgo, implica un estudio detallado de los parámetros de las pruebas, así como, un estudio del funcionamiento diferencial de los ítems (Arias, 2008; Hambleton,

[4] Para una mayor profundización puede verse: Csapó, Ainley, Bennett, Latour y Law (2012), capítulo 4 "*Technological Issues for Computer-Based Assessment*" (p.143-230), que recoge cuesiones como la validez de las pruebas según el formato, normas de seguridad ene las pruebas, así como comparabilidad.

Clauser, Mazor y Jones, 1993). Unido a la realización de un estudio completo de la validez para garantizar el correcto estudio de la equivalencia entre ambos formatos de prueba (papel y online).

CAPÍTULO 3: Propuesta Metodológica: Funcionamiento Diferencial de Versiones

"Toda educación auténtica se transforma en investigación del pensar".

(Freire, P, 1999, p.30)

3.1. Aportación Metodológica

Una de las mayores preocupaciones cuando informatizamos un test convencional es saber si se tendrá la posibilidad de comparar ambas versiones para determinar si el modo de presentación de los test origina puntuaciones equivalentes, o si, por el contrario, es necesario cambiar el baremo de referencia dado que la diferencia de formato afecta al rendimiento de los sujetos.

Cuando revisamos la literatura especializada nos encontramos principalmente con estudios centrados en la comparabilidad de las puntuaciones y de las características psicométricas de ambas versiones. Concretamente, se estudian las medias, desviaciones típicas, consistencia interna y validez de ambas versiones, para establecer confirmaciones sobre la comparabilidad y, por tanto, la equivalencia psicométrica entre ambas versiones (Boo y Vispoel, 2012; Finger y Ones, 1999).

Existe un grupo de autores partidarios de considerar que el modo en que se conduce una prueba da lugar a diferencias significativas en el rendimiento, a favor de los sujetos que realizan la prueba en papel. Estos estudios concluyen que la versión informatizada resulta más difícil que la versión en papel (Applegate, 1993; Bennett, Braswell, Oranje, Sandene, Kaplan y Yan, 2008; Cerillo y Davis, 2004, citados en Paek, 2005; Ito y Sykes, 2004; Karkeen, Kim y Fatica, 2010; Jackel, 2014; Russell, 1999; Sandene, Horkay, Bennett, Allen, Braswell, Kaplan y Oranje, 2005; Way, Davis y Fitzpatrick, 2006).

En menor medida, hay algunos estudios que confirman que la puntuación en la versión informatizada es superior a la de la versión en papel y lápiz (Chin, Donn y Conry, 1991; Greaud y Green, 1986, citados en Arribas, 2004; Choi y Tinkler, 2002; y Kalogeropoulos, Tzigounakis, Pavlatou y Boudouvis, 2013).

Worrell, Duffy, Brady, Dukes y Gonzalez-DeHass (2015) concluyen, en un estudio sobre la comprensión lectora, que los sujetos que realizaron la prueba por ordenador obtuvieron mejores resultados que los que la hicieron en soporte papel.

Vispoel, Rockiln y Wang (1994) destacan, tras revisar diferentes estudios, que hay una preferencia hacia los test informatizados (en concreto hacia los adaptativos) sobre los test en papel, siempre y cuando los primeros ofrezcan las mismas condiciones que los test convencionales (posibilidad de revisión de los ítems, omisión, etc.).

En contraposición, son numerosos los estudios en los que el modo de aplicación del test no produce diferencias estadísticamente significativas: Al-Amri (2007); Arce-Ferrer y Guzmán (2009); Čandrlić, Ašenbrener y Holenko (2014); Csapó, Molnár y Nagy (2014); Ita, Kecskemety, Ashley y Morin (2014); Higgins, Gray, Symeonidis y Tsintsifas (2005); Horkay, Bennett, Allen, Kaplan y Yan (2006); Ito y Sykes (2004); Johnson y Green (2006); Keng, McClarty, y Davis (2008); Kim y Huynh (2007); Lottridge, Nicewander y Mitzel (2011); Moon (2013); Oregon Department of Education (2007); Paek (2005); Pearson (2002–2003); Poggio, Glasnapp, Yang y Poggio (2005); Pommerich (2004); Russell y Haney (1997); Wang (2004); Wang, Jiao, Young, Brooks y Olson (2007-2008); Yu, Livingston, Larkin y Bonett (2004).

En un estudio más reciente, Laurie, Bridglall y Arseneaukt (2015) señalan que no hay diferencias significativas entre las puntuaciones medias en ambas versiones, pero sí registran diferencias significativas a favor de la prueba en papel en lo concerniente a los criterios de sintaxis y favorables a la prueba online en lo relativo a la ortografía.

En la Universidad Autónoma de Baja California se aplicó un Sistema Informatizado de Exámenes (SICODEX), primero en versión papel y más tarde por ordenador, vía web. Se evaluaron habilidades y conocimientos básicos como prueba de admisión para entrar en la Universidad (EXHCOBA), y estas evaluaciones se llevaron a cabo en diferentes universidades (México, EEUU y Canadá) entre 1993 y 1995. La versión informatizada obtuvo apropiados niveles de fiabilidad y ambas versiones se consideraron equivalentes, no mostrando diferencias significativas entre ellas (Backhoff, Ibarra, Rosas y Larrazolo, 1999).

A través de un análisis multinivel, en la investigación llevada a cabo por Moon (2013) se volvió a demostrar que los resultados en una prueba de matemáticas no presentaban diferencias estadísticamente significativas según el modo de aplicación.

También existen estudios centrados en evaluar la comparabilidad de las puntuaciones obtenidas en dos versiones de una misma prueba, llevando a cabo estudios sobre sus características psicométricas (estudiando las medias, desviaciones típicas, consistencia interna y validez de ambas versiones). Un ejemplo de ello son los estudios realizados por Boo y Vispoel (2012) y Finger y Ones (1999), que confirman la comparabilidad y por tanto la equivalencia psicométrica entre ambas versiones.

Ante esta variedad de resultados, es importante matizar que hay características a valorar que pueden producir diferencias en el rendimiento, ocasionando un funcionamiento diferencial de las versiones.

Podríamos englobar estas características en dos grandes grupos: un grupo centrado en las características de los participantes y otro centrado en las características de la pruebas.

Centrados en las características de los participantes

Algunos ejemplos de este tipo de características pueden ser el género, el nivel socioeconómico, la ansiedad o la familiaridad con los ordenadores, todas ellas variables interesantes a la hora de evaluar, ya que pueden repercutir en el rendimiento de los individuos.

A continuación haremos un recorrido sobre algunos de los estudios que han trabajado con estas variables y presentaremos sus conclusiones.

- *Nivel Socioeconómico*

Cuando se habla del nivel socioeconómico se hace referencia al conjunto de variables relacionadas con el entorno social y económico del sujeto.

Es conocida la influencia que este tipo de variables tienen en el rendimiento educativo. Remontándonos al informe Coleman (Kain y Singleton, 1996), y al estudio de Cardona, González y Gutiérrez (1973), e incluso a estudios relativamente pioneros (Loevinger, 1940; Maller, 1933; Nelf, 1938; Thorndike y Woodward, 1942), vemos que

se registran diferencias en el rendimiento, siendo éstas ascendentes en la medida en que aumenta el estrato socioeconómico.

En la actualidad son muchos los estudios al respecto que ofrecen las mismas conclusiones (Armenta, Pacheco y Pineda, 2008; Caro, McDonald y Willms, 2009; Ferrera, Cebada y Chaparro, 2013; Gil-Flores 2011; Olmedo, 2007; Rumberger, 2004; Van Ewijk y Sleegers, 2010). Es tal la importancia que toma dicha variable que las evaluaciones internacionales tienen una mención especial al respecto (OCDE, 2013-PISA 2012).

Podemos destacar estudios como el de MacCann (2005), llevado a cabo en estudiantes de entre 15 y 16 años en New South Wales (Australia), donde se aplicó una prueba en versión informatizada y otra en papel, con la intención de determinar si el género, el nivel socioeconómico y/o el tipo de ítem interactuaban con el modo en que se aplicaron las pruebas. Los resultados indicaron diferencias estadísticamente significativas entre el nivel socioeconómico y el modo de aplicación: para los estudiantes con bajo nivel socioeconómico, la media en la prueba online era un 1% inferior a la media obtenida en la prueba en papel; mientras que en estudiantes con un nivel socioeconómico alto, las medias tendían a igualarse.

Pero, en contraposición, en un estudio llevado a cabo por Sanden, et al. (2005) no se encontraron diferencias significativas entre el rendimiento y el nivel educativo de los padres.

- *Género y etnia*

Son muchos los estudios que investigan la influencia que el género y la raza puedan tener en el rendimiento de los estudiantes atendiendo al modo de aplicación de la prueba.

Ejemplo de ello son los estudios llevados a cabo por Bennett et al. (2008), Clariana y Wallace (2002), Sim y Horton (2005), Sandene et al. (2005), y MacCann

(2005), que encuentran que no hay diferencias entre el modo de administración de las pruebas y el género.

Otros estudios, por el contrario, demuestran la existencia de diferencias en función del género. En general, el impacto es mayor para las mujeres en los test informatizados. Siguiendo esta idea, Segall (1997) determina que las mujeres examinadas obtuvieron mejores resultados en la prueba en papel que en la informatizada.

Gallagher, Bridgeman y Cahalan (2002) exponen que, cuando las diferencias de género se comparan con los grupos raciales, las diferencias estadísticamente significativas se encuentran sólo para el grupo de los examinados blancos. El impacto mayor (aunque con diferencias pequeñas) se produce en la versión informatizada para las mujeres blancas. Llevaron a cabo una revisión de diferentes pruebas informatizadas que fueron analizadas en función de la etnia, los resultados demostraron que los afroamericanos e hispanos examinados obtuvieron, en algunos casos, mejores resultados que en la versión en papel.

También Pino, de la Iglesia, Gialluca y Weiss (1980) destacan estas diferencias, pero las atribuyen a un aumento en la motivación de los afroamericanos examinados durante la realización de prueba informatizada (Gallagher, Bridgeman y Cahalan, 2002).

Oltman (1994) estudió los efectos en el rendimiento en Lectura y Matemáticas producidos por la manipulación del ratón en función de su grado de complejidad, del más simple (en el que sólo requería un clic para contestar) al más complejo (requería dos clic para contestar). Dicho estudio fue aplicado a un grupo de minorías (hispanos, nativos americanos, afroamericanos) y a otro grupo de estudiantes universitarios blancos. Los análisis revelaron una interacción significativa entre el grupo étnico y el tipo de tarea (pero no tan pronunciada como para que esta interacción fuera digna de tomarse en consideración): las minorías necesitaron más tiempo para contestar y consiguieron calificaciones más bajas que los estudiantes blancos.

- *Conocimiento informático o familiaridad*

Es posible encontrarnos con sesgo de método, es decir, con un error en la medida que evita la comparación entre los grupos debido principalmente a la desigualdad o falta de homogeneidad entre las muestras, o, como en nuestro caso, cuando el modo de aplicación de la prueba es diferente en ambos grupos. Esto puede originar problemas en las respuestas ofrecidas por los sujetos, debidas a la falta de familiaridad con el formato de la prueba y los ítems (Van de Vijver y Poortinga, 2005; Van de Vijver y Tanzer, 2004).

En lo relativo a los conocimientos informáticos de los alumnos, debido al uso que hacen de los ordenadores en el hogar y en el centro educativo, Bennett, et. al (2008), Goldberg y Pedulla (2002) y Pomplun y Custer (2005) defienden la idea de que el conocimiento informático es un predictor significativo del rendimiento de los examinados, lo que ocasiona que los estudiantes que tienen menor experiencia con los ordenadores obtengan puntuaciones más bajas.

Bennett et al. (2008) y Johnson y Green (2006) identificaron dos factores que generan diferencias en el rendimiento de los sujetos en función del modo de aplicación de la prueba. Uno de ellos es la familiaridad con ordenadores, estos estudios determinaron que los sujetos con experiencia en el manejo de ordenadores (por ejemplo, porque hayan realizado pruebas con anterioridad) obtienen un mejor rendimiento. El otro factor es la viabilidad necesaria para convertir la versión en papel de un ítem de respuesta construida a la versión informatizada.

Por otro lado, Taylor, Jamieson, Eignor y Kirsch (1998) llevaron a cabo un estudio en el que tenían en cuenta la familiaridad de los sujetos con el equipo informático y los resultados conseguidos en pruebas TOEFL, y aunque se produjeron diferencias en el rendimiento en función de la familiaridad con la prueba, se concluyó que la familiaridad con el ordenador no jugaba un papel relevante en el rendimiento en la prueba TOEFL.

Schade, Hernández y Elgueta (2005) estudiaron la aplicabilidad del instrumento en modo informatizado mediante la implementación informatizada de un test convencional para evaluar la memoria MEMOPOC. Los resultados evidenciaron puntos fuertes, como el aumento en la motivación, menor agotamiento y mayor atractivo ante estas versiones, y puntos débiles, como la falta de comprensión de las instrucciones.

Otro factor importante, asociado a la falta de familiaridad con los ordenadores, es la ansiedad. Wheadon y Adam (2007) abordan las ideas de Kveton, Jelinek, Voboril y Klimusova (2007) y Smith y Caputi (2007), en las que observaron que la poca familiaridad con los medios informáticos provoca ansiedad y, por consiguiente, afecta negativamente en los resultados.

Cuando la familiaridad con el ordenador es mayor, se observa que los estudiantes se sienten más motivados y prefieren realizar la prueba informatizada (Rajmil, Robles, Murillo, Rodríguez-Arjona, Azuara, Ballester y Codina, 2015). En estudios en los que se ha tenido en cuenta la motivación de los estudiantes ante las pruebas, los resultados nos informan de que los efectos producidos por la motivación no son tan grandes como los efectos producidos por el modo de administración (Kiplinger y Linn, 1996 y O'Neill, Sugrue y Baker, 1996).

Muchos estudios abordan una característica a considerar al realizar pruebas informatizadas: la fatiga o el cansancio que puede provocar la lectura constante en la pantalla del ordenador, así como la velocidad lectora (Solak, 2014).

El reciente estudio llevado a cabo por Karay, Schauber, Stosch y Schuttpelz-Brauns (2015) en la Universidad de Berlín evaluó a 266 estudiantes de medicina (papel = 132, ordenador = 134) para ver si existían diferencias en el rendimiento en función del modo de prueba (aplicada en papel o en ordenador). Los resultados reflejaron la no existencia de diferencias estadísticamente significativas entre el modo de aplicar la prueba y el rendimiento, pero sí se detectó menor tiempo para completar la prueba en los sujetos que la hicieron por ordenador.

Mead y Drasgow (1993) observaron pocas diferencias en los parámetros de los ítems en test de potencia y una mayor diferencia en test de rapidez (Puhan, Boughton, Kim, 2007 y Sawaki, 2001).

Russell (1999) realizó un estudio centrado en pruebas con preguntas de respuesta abierta, en papel e informatizadas, en las áreas de matemáticas, lenguaje y ciencias:

> Para las dos versiones, en lenguaje no se encontró ningún efecto significativo. Sin embargo, para los estudiantes cuya velocidad con el teclado es de al menos 0,5 o de la mitad de una desviación estándar por encima de la media, la realización de la prueba de lengua y literatura por ordenador tuvo un efecto positivo moderado. Por el contrario, para los estudiantes cuya velocidad con el teclado fue de 0,5 de desviación estándar por debajo de la media, la realización de las pruebas en el ordenador tuvo un efecto negativo.
>
> Para el test de matemáticas, la realización de la prueba por ordenador tuvo un efecto negativo, pero este efecto fue menos pronunciado en aquellos sujetos que presentaban una velocidad alta con el teclado.
>
> Para la prueba de ciencias, el rendimiento en la versión informatizada tuvo un efecto positivo en el grupo.

En contraposición, Sawaki (2001), recoge otros estudios como el de Neuman y Baydoun en 1998, en el que se consideran equivalentes unas pruebas de rapidez llevadas a cabo en formato papel y las informatizada. Las investigaciones llevadas a cabo por Fish y Feldmann en 1987; Feldman y Fish en 1988, y Yessis en el 2000 no detectaron diferencias significativas en la velocidad lectora según los diferentes modos de aplicación.

Estudios como los de Chen (2015) y Chen y Catrambone (2015) demuestran en una prueba de comprensión lectora, en papel y online, que los participantes invierten más tiempo en la lectura en papel que en pantalla; aun así, los resultados indican que las diferencias en los niveles de comprensión no son significativas.

De nuevo, en el estudio de Sun, Shieh y Huang (2013) no se aprecian diferencias estadísticamente significativas en la comprensión lectora en función del modo de presentación del texto, es decir, en pantalla de ordenador o papel.

Características relativas al propio test

Se trata de características propias del test, fijadas de antemano, para elaborar la prueba. Estas características pueden tener consecuencias en el rendimiento de los sujetos. Un ejemplo de ello puede ser la fuente utilizada en ambas versiones, o el orden de las preguntas y de las alternativas: cualquier cambio en estas características puede afectar en el rendimiento. Concretamente, la calidad de la presentación en la pantalla del ordenador (tamaño y resolución) puede influir negativamente en el rendimiento (Schenkman, Fukada, y Perrson, 1999), así como igualmente pueden influir de forma negativa problemas derivados de la caída de la red en el momento de la aplicación (Kingston, 2002).

Wang y Shin (2009) recogen algunos estudios que han examinado la relación entre el modo de administración y características de los ordenadores, tales como tamaño de pantalla, resolución y tamaño de fuente. Por ejemplo, McKee y Levinson (1990) indican que estos factores pueden cambiar la naturaleza de una tarea de manera drástica, hasta el punto de que los ítems administrados en ordenador y en papel no midan el mismo constructo.

- *Flexibilidad*

Las dos formas de administración serán equivalentes siempre que se garanticen las mismas condiciones de flexibilidad. Lunz en 1995 y Vispoel en 2000 han trabajado con estas características, pero sin obtener resultados del todo claros (Revuelta, Ximénez y Olea, 2003). La flexibilidad con la que los examinados pueden interactuar con los ordenadores es otra posible explicación para las diferencias de rendimiento observadas en los diferentes modos de administración.

Por ello, es importante garantizar la posibilidad de revisión de los ítems en las pruebas informatizadas, al igual que sucede en las pruebas convencionales. Sykes e Ito (1997) aseguran que los test informatizados en los que no se permite la revisión de los ítems resultan frustrantes, producen ansiedad y, por ende, producen resultados inferiores a los de los test en papel. Esta idea es compartida por otros autores como Vispoel, Wang, de la Torre, Bleiler y Dings (1992), que declararon que permitiendo la revisión de los ítems aumentaron significativamente el número de respuestas correctas.

En el estudio llevado a cabo por Olea, Revuelta, Ximénez y Abad en 2000, los autores apreciaron que en la versión informatizada, donde se permite la revisión de los ítems, se dan menores niveles de ansiedad y mayores tasas de respuestas cambiadas, además de un aumento del número de respuestas correctas y un incremento del nivel de rasgo medio estimado. Sin embargo, hay estudios que obtienen resultados diferentes, como el de Vispoel en 2000, que observa cómo en las pruebas informatizadas en las que es permitida la revisión de ítems, el número de respuestas cambiadas es similar al de los test en papel, pero el porcentaje de sujetos que modificaron sus respuestas es menor: entre ellos, más del 50% aumentó su puntuación en la prueba (trabajos citados en Revuelta, Ximénez y Olea, 2003).

- *Tipo de ítem*

Son muchos los estudios que investigan la influencia que pueda tener el tipo de ítem de una prueba (abierto, cerrado, opción múltiple, etc.) según el soporte en que sea aplicado (papel o informatizado).

Russell y Haney (1997) recogen un estudio llevado a cabo con estudiantes de Secundaria para ver el efecto que tiene el modo de administración del test (papel y lápiz o informatizado) en el rendimiento de los sujetos, utilizando preguntas de opción múltiple y abiertas. Los resultados indican que las respuestas a las preguntas de opción múltiple, según el modo de administración de las pruebas, no difieren de forma significativa; en cambio, en las preguntas abiertas, aquellos sujetos que están acostumbrados a escribir en ordenador obtienen puntuaciones mayores que los que suelen escribir a mano.

Clariana y Wallace (2002) y Wang y Shin (2009) mencionan a otros autores que no comparten estos resultados, como Mazzeo, Druesne, Raffeld, Checketts, y Muhlstein en 1991, o el estudio de Taylor, Kirsch, y Eignor en 1999. El problema principal es que el conocimiento es medido y definido de muchas y muy diferentes formas, por lo que es necesario garantizar la validez de estas habilidades informáticas

En lo que respecta a la calidad escrita de los ítems abiertos, Grejda (1992) no encontró diferencias entre la calidad del texto y la calidad de la escritura en los dos formatos. Esta idea es compartida por Nichols (1996), aunque éste añade que los estudiantes que realizaron la prueba a ordenador pasaron más tiempo escribiendo. Owston (1991) destaca que las redacciones creadas en el ordenador se valoraron significativamente mejor que las producidas en papel.

Revisadas las aportaciones de los últimos años, y siguiendo la idea de Rowan (2010), en nuestro trabajo no solo se estudia la equivalencia mediante la evaluación de las distribuciones de las puntuaciones, también se utiliza una propuesta metodológica inspirada en el Funcionamiento Diferencial de los Ítems y en el Funcionamiento Diferencial de los Sujetos, que tomará por nombre *"Funcionamiento Diferencial de las Versiones" (Differential Version Functioning).*

3.1.1. Propuesta Metodológica

Los desarrollos de la Teoría de Respuesta al Ítem (TRI), junto con los avances técnicos y de software, hacen posible la aplicación de modelos psicométricos cada vez más complejos.

Una de las aplicaciones de la TRI, en la que se enmarca este trabajo, son los estudios del Funcionamiento Diferencial de los Ítems. En términos generales, estos trabajos se centran en comprobar si la medición de los test es objetiva y si las posibles diferencias encontradas con respecto a un rasgo son reales o producto del test.

3.1.1.1. Funcionamiento Diferencial de los ítems

Los estudios sobre el Funcionamiento Diferencial de los Ítems[5] (de ahora en adelante DIF) pueden remontarse a las pruebas de inteligencia llevadas a cabo por Binet y Simon (1916); en estas pruebas fue necesaria la eliminación de algunos ítems por la existencia de variabilidad en el rendimiento debida a factores culturales.

En los años 50, Eells, Davis, Havighurst, Herrick y Tyler (1951) llevaron a cabo un estudio pionero sobre el sesgo de los ítems. En este estudio, realizado en la Universidad de Chicago, los autores demostraron empíricamente cómo algunos ítems se comportaban de forma diferente según el nivel socioeconómico de los participantes y según las características de los ítems (formato y contenido). Aquellos grupos de sujetos de clase social más baja presentaban complicaciones a la hora de entender ciertos términos con los que no estaban familiarizados, lo que ocasionaba una ventaja para el grupo de sujetos de nivel cultural más alto. Así, los autores mostraron la existencia de sesgo cultural en algunos ítems, propiciando un gran interés entre los investigadores por dar respuestas a los motivos de dichas diferencias.

Tras el impacto social y político de este tipo de trabajos, en Estados Unidos, durante los años 70, se produjo una corriente orientada a la defensa de los derechos civiles, promovida por las diferencias que se apreciaban en los test, fruto de las características étnicas, socioeconómicas y culturales. Estas diferencias desfavorecían a algunos colectivos, dado que los test eran utilizados para la selección de personal, tanto en el terreno educativo como en el laboral; por este motivo surgieron protestas sociales para garantizar la defensa por la igualdad de los derechos.

Es en este momento en el que la palabra “sesgo” comienza a asociarse a otros términos como imparcialidad, injusticia, etc., es decir, a adquirir connotaciones

[5] Acrónimo, en inglés, del Differential Item Functioning.

negativas, puesto que se perjudicaba a los grupos con condiciones menos favorables (Cole, 1993).

Autores como Fidalgo (1996) y Cole (1993) consideran que un instrumento correcto de medida tiene que ser neutral, y en ningún caso favorecer a un grupo concreto, es decir, no tiene que discriminar; de ser así, el test está sesgado. A este respecto, Muñiz (1992) indica:

> *"el problema del sesgo viene acompañado de serias implicaciones sociales en el uso de los test, pues de darse tal sesgo ciertos grupos sociales, clásicamente blancos-negros, mujeres-hombres, pobres-ricos, etc., cualquiera otra partición es posible, sufrirán la consecuencia"* (p.198).

Por todo lo mencionado en líneas anteriores, es el momento de distinguir la asociación que hasta el momento se producía entre las diferencias reales en el test (de contenido técnico) y las de tipo cultural (de contenido social). En la década de los 70 y los 80 se elaboraron y aplicaron técnicas y métodos para el análisis y estudio del sesgo, además de una profundización en el concepto y acepciones del término, incorporándose en el Educational Testing Service (American Educational Research Association, 1986).

Al adaptar un test es posible que se produzcan errores, conocidos como sesgos. El sesgo es una fuente de invalidez, que distorsiona de forma sistemática los resultados o puntuaciones del test en los grupos de sujetos estudiados, causando un error sistemático (Ackerman, 1992). Para Camilli y Shepard (1994) el estudio del sesgo supone investigar los motivos por los que determinados ítems, en función de algunas variables, presentan funcionamiento diferencial. Es la distorsión sistemática de los resultados la que provoca la invalidez del test.

Son muchos los autores que han abordado el tema; un ejemplo de ello es el trabajo de Chahín - Pinzón (2014), quien realiza la siguiente clasificación: sesgo de constructo, sesgo de método y sesgo de los ítems.

Sesgo de constructo: hace referencia a la validez de las medidas del mismo constructo en distintos grupos. Atendiendo a las ideas recogidas por Van de Vijver y Tanzer (2004), Chahín - Pinzón (2014), Fernández, Pérez, y Alderete, Richaud y Fernández Liporace (2011), el sesgo de constructo deriva de la diferencia existente en la medición del constructo; es decir, el constructo medido no es el mismo en ambos grupos comparados. Este sesgo se produce principalmente cuando lo que se pretende es adaptar un test a contextos culturales diferentes, lo que puede provocar variaciones en el significado del propio constructo medido, así como diferencias en los resultados causadas por la disparidad cultural de los sujetos.

Sesgo de método: se trata de un error en la medida que evita la comparación entre los grupos, debido principalmente a la desigualdad o falta de homogeneidad entre las muestras; o cuando el modo de aplicación de la prueba es diferente en ambos grupos. Es éste un aspecto clave de nuestro trabajo, dado que comparamos dos pruebas administradas en dos formatos diferentes, papel y lápiz e informatizado, hecho que puede originar problemas en las respuestas ofrecidas por los sujetos, debidos a la falta de familiaridad con el formato en que se aplica la prueba y los ítems (Van de Vijver y Poortinga, 2005; Van de Vijver y Tanzer, 2004).

Sesgo de los ítems: son los errores que tienen los ítems porque actúan injustamente en alguno de los grupos, con respecto al rasgo latente medido. Esto es debido a que algunos ítems miden otros rasgos ajenos al rasgo medido, lo que implica un atentado a la validez del test (Gómez e Hidalgo, 1997). Asimismo, hay que señalar que un ítem presenta sesgo si los grupos que se comparan muestran la misma puntuación en el test, pero las puntuaciones medias de los grupos y los índices de dificultad para los grupos comparados son estadísticamente distintos. Atendiendo a las ideas de Hambleton y Zenisky (2011), cuando llevamos a cabo una adaptación es imprescindible atender y verificar que los ítems tengan similares características psicométricas (como el nivel de dificultad y de discriminación). Estos errores pueden deberse a traducciones incorrectas, connotaciones diferentes que deriven en palabras distintas en ambos test, y al desconocimiento del contenido del ítem (Hambleton, 2005; Muñiz y Hambleton, 1996; Sireci, 2011; Van de Vijver y Leung, 2000; Van de Vijver y Poortinga, 2005).

Para llevar a cabo la detección del sesgo se utilizan algunas aproximaciones estadísticas, aplicando criterios tanto externos como criterios internos al test. Los criterios externos para la detección del sesgo externo son los errores que se cometen en las puntuaciones debido a la relación con variables ajenas al test (características culturales, económicas, étnicas,…), mientras que los criterios internos para la detección del sesgo interno hacen referencia fundamentalmente a las puntuaciones alcanzadas en el test, es decir, a las propiedades psicométricas de los ítems del test. El sesgo interno es un error sistemático que se produce a lo largo de todo el proceso de medida, sin atender a otras características externas de carácter sociocultural (Martínez, 1997; Whitmore y Schumacker, 1999).

Ante esta controversia, desde la psicometría se comenzó a analizar el sesgo como un problema propio del instrumento, es decir, de las características psicométricas del test, y no como un aspecto discriminativo fruto de algunas características de los grupos (Fidalgo, 1996). Fue entonces cuando Holland y Thayer (1988) sustituyeron el término "sesgo" por la expresión "Funcionamiento Diferencial de los Ítems", más preciso, estudiando ítems que tienen funcionamiento distinto en sujetos o grupos con habilidades parecidas (Gómez e Hidalgo, 1997).

Atendiendo a estas ideas, podemos declarar que los índices estadísticos para el análisis del DIF no demuestran ni dan evidencia del sesgo de los ítems de la prueba (Camilli y Shepard, 1994; Muñiz, 1997). Es por ello, por lo que ambos términos tienen que estar bien diferenciados; a pesar de estar relacionados no podemos considerarlos como equivalentes. Hay una clara diferencia entre ambos términos, puesto que cuando hablamos de DIF lo hacemos desde un punto de vista metodológico o estadístico, mientras que cuando nos referimos al sesgo hablamos desde una perspectiva teórica (Camilli y Shepard, 1994; Fidalgo, 1996; Muñiz, 1997; Shealy y Stout, 1993).

Es importante señalar que cuando un ítem presenta DIF no necesariamente requiere un sesgo en el ítem. Es imprescindible acudir a expertos que valoren cuidadosamente el contenido del ítem y determinen si contiene lenguaje, términos o símbolos que beneficien o perjudiquen a uno de los grupos (Allalouf, Hambleton y

Siresi, 1999; Chahín-Pinzón, 2014; Fernández, Pérez, Alderete, Richaud y Fernández Liporace, 2011; Van de Vijver y Tanzer, 2004; Zumbo, 1999). De ser así, entonces sí podemos considerar que un ítem está sesgado.

Fidalgo y Ferreres (2002), además de indicar la importancia que tiene detectar el DIF en los ítems del test, indican la relevancia que supone verificar el impacto del DIF; es decir, estudiar e identificar cuáles son las posibles razones o causas que ocasionan el DIF en un ítem o en el test, que hacen que este no sea ecuánime para con los grupos estudiados (Gómez y Navas, 1998).

El que las propiedades psicométricas de un ítem en ambos grupos sean diferentes nos revela que dicho ítem presenta DIF, pero el que un grupo sistemáticamente obtenga resultados inferiores no tiene por qué deberse al DIF, ya que puede deberse a diferencias reales, conocidas como impacto o diferencias válidas (Millsap y Everson, 1993), o a diferencias artificiales, propias del test, conocidas como DIF (Andrich y Hagquist, 2012; Camilli y Shepard, 1994; Gómez, Hidalgo y Guilera, 2010; Van de Vijver y Leung, 1997).

En resumen, cuando hablamos de impacto hacemos referencia a las diferencias reales que existen entre los grupos de comparación, y, por tanto, la probabilidad de responder correctamente a un ítem será siempre mayor en un grupo que en otro, puesto que existen diferencias en la habilidad medida que favorece a un grupo, causadas por una diferencia real en la variable. Por el contrario, cuando hablamos de DIF se asume que ambos grupos parten del mismo nivel de habilidad medida, y por tanto que las posibles diferencias que nos encontremos entre los grupos se deban a problemas de validez propios del test, y no de los sujetos (Boone, Staver y Yale, 2014; Cuevas, 2013; Elosua, 2006; Herrera, 2005; Holland y Wainer, 2012).

El funcionamiento diferencial de los ítems estudia aquellos ítems que producen diferentes resultados en sujetos con habilidades similares, por lo que se considera que un ítem tiene DIF si presenta un funcionamiento diferente en varios grupos; esto es, que haya una probabilidad distinta de responder correctamente a los ítems entre los grupos

comparables o igualmente capaces (Angoff, 1993; Cuevas, 2013; Elosua, 2006; Elosua y López-Jáuregui, 2007; Hambleton, Swaminathan y Rogers, 1991; Herrera, 2005).

Según Gómez, Hidalgo y Guilera (2010), *"un ítem presenta DIF cuando grupos igualmente capaces presentan una probabilidad distinta de responderlo con éxito o en una determinada dirección en función del grupo al que pertenecen"* (p.76).

Cuando hablamos de DIF hacemos referencia a grupos que son comparables. Normalmente se tiende a comparar dos grupos, siendo el grupo mayoritario conocido como Grupo de Referencia (R) y que el grupo minoritario como Grupo Focal (F) (Gierl y Khaliq, 2001).

Desde la perspectiva enfocada a la Teoría de Respuesta al Ítem, un ítem presenta DIF cuando se observan las diferencias entre las puntuaciones del grupo de referencia y el grupo focal con la misma variable latente, por medio de las Curvas Características de los Ítems (Kim y Cohen, 1998; Mazor, Hambleton y Clauser, 1998).

Un ejemplo de ello, puede verse en la figura 1. En este caso, estamos ante un ítem que no presenta DIF. Esto sucede porque las Curvas Características de los Ítems (de ahora en adelante CCIs[6]) del grupo de referencia (en nuestro caso los sujetos que realizan la prueba en papel) y del grupo focal (sujetos que hacen la prueba online), coinciden exactamente. Es decir, tanto los parámetros de dificultad como de discriminación en ambos grupos son idénticos.

[6] Curva Característica del Ítem, definida por Camilli y Shepard (1994, p.47) como *"la función que relaciona la probabilidad de responder correctamente a un ítem con la habilidad medida por el instrumento que contiene a dicho ítem"*.

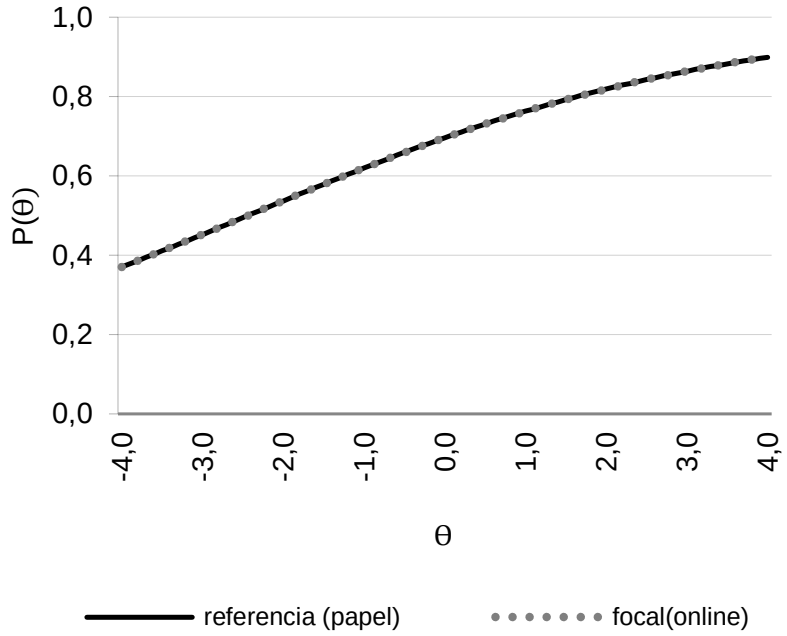


Figura 1. Representación gráfica (CCI) de la no presencia de DIF

Todo lo contrario sucede con un ítem que presenta DIF: los parámetros de dificultad y de discriminación en ambos grupos son distintos. Por ello, las CCIs no coinciden en todos los niveles del atributo.

Podemos encontrarnos varios tipos de DIF:

- *DIF Uniforme o consistente:*

Sucede cuando las CCIs son diferentes en los grupos comparados (grupo de referencia - R y grupo Focal - F) y, por lo tanto, las curvas no se cruzan en ningún nivel de la variable medida porque existe un grupo aventajado. Los parámetros de dificultad en las dos CCIs son diferentes, mientras que los parámetros de discriminación son iguales, lo que hace que las dos CCIs sean paralelas (Gómez, Hidalgo y Guilera, 2010; Gómez e Hidalgo, 1997; Rogers y Swaminathan, 1993; Roussos, Schnipke y Pashley, 1999).

Podemos observar gráficamente un ejemplo de DIF uniforme o consistente en la figura 2.

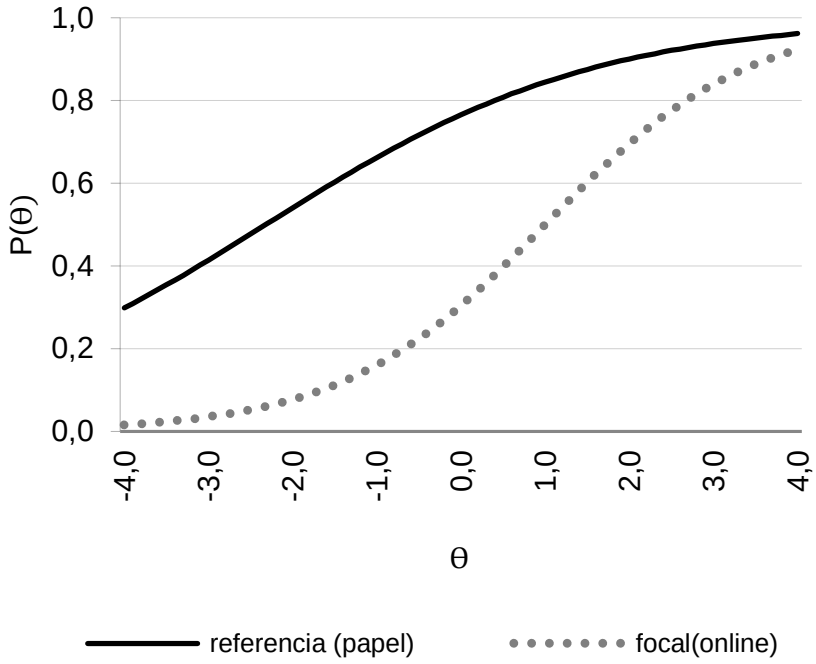


Figura 2. Representación gráfica (CCI) de DIF uniforme o consistente
Nota: Grupo de Referencia con θ =1 tiene una probabilidad de éxito de 0,5.
Grupo Focal con θ =1 tiene una probabilidad de éxito de 0,6.

Podemos apreciar como para un mismo nivel de habilidad (θ) los valores de las probabilidades de éxitos P(θ) no son los mismos, sino que siempre son superiores para uno de los dos grupos. En este caso, podemos hablar de DIF desfavorable para los estudiantes que hacen la prueba online (Grupo Focal), puesto que los que realizan la prueba en papel (Grupo Referencia) siempre obtienen mejores resultados.

- *DIF No uniforme o inconsistente:*

Se produce cuando las probabilidades de éxito no son las mismas. No se trata de una tendencia constante, sino que en un momento dado favorece a un grupo y en otro momento a otro, invirtiéndose la tendencia (Rogers y Swaminathan, 1993). Podemos observar gráficamente un ejemplo de DIF no uniforme o inconsistente en la figura 3.

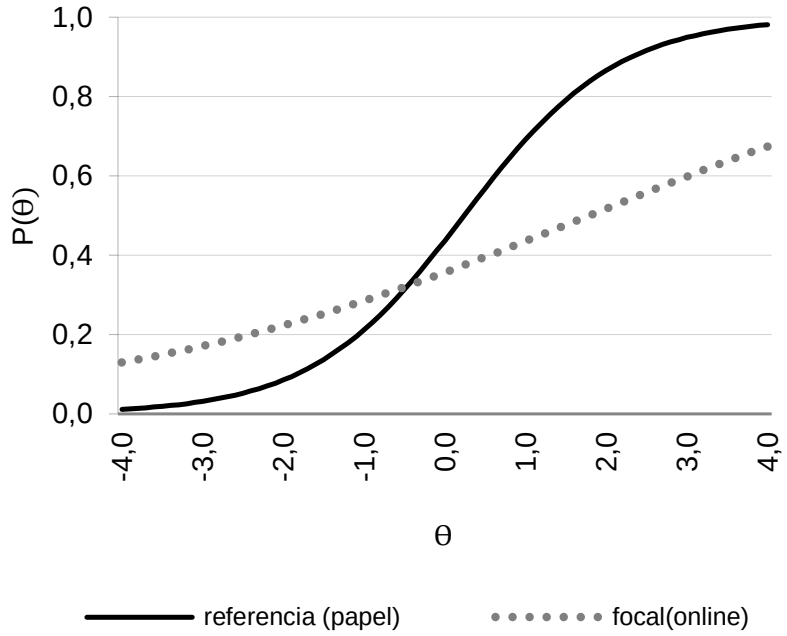


Figura 3. Representación gráfica (CCI) de DIF no uniforme

Como se puede apreciar en el ejemplo, el Grupo de Referencia (prueba papel) con θ = -2,0, tiene una probabilidad de éxito de 0,1; mientras que el Grupo Focal, con el mismo valor de θ = -2,0, tiene una probabilidad de éxito de 0,2. Lo que representa que el ítem es algo más difícil para el Grupo de Referencia, ya que $P_F(\theta) > P_R(\theta)$.

Todo lo contrario sucede cuando observamos en el ejemplo cómo el grupo de Referencia (prueba papel) tiene una probabilidad de éxito de 0,9 con θ =2,0 y el grupo Focal de 0,5 con el mismo nivel de habilidad (θ). En este caso se aprecia que el ítem es algo más difícil para el Grupo Focal, ya que $P_R(\theta) > P_F(\theta)$.

Swaminathan y Rogers (1990) realizan una doble clasificación del DIF no uniforme, dando lugar al DIF no uniforme simétrico y el DIF no uniforme mixto.

- *DIF no uniforme Simétrico*

Se trata de un tipo de DIF que se produce cuando la interacción se lleva a cabo en el centro de las CCIs, por lo que manifiesta que en ambos grupos el parámetro de dificultad es constante y el de discriminación diferente. En la figura 4 podemos observar gráficamente un ejemplo de DIF no uniforme sistemático.

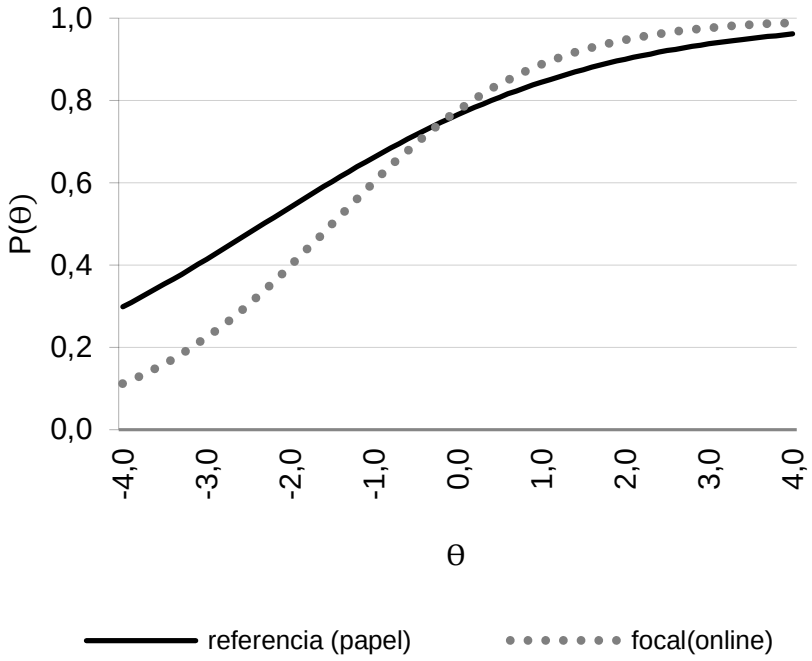


Figura 4. Representación gráfica (CCI) de DIF no uniforme (simétrico)

- *DIF no uniforme Mixto*

También conocido como asimétrico (Hidalgo, Gómez y Padilla, 2005), es un tipo de DIF que se produce cuando la interacción de las CCIs es asimétrica; es decir, cuando los parámetros de dificultad y discriminación en ambos grupos son distintos. En la figura 5 se muestra un ejemplo gráfico de DIF no uniforme mixto.

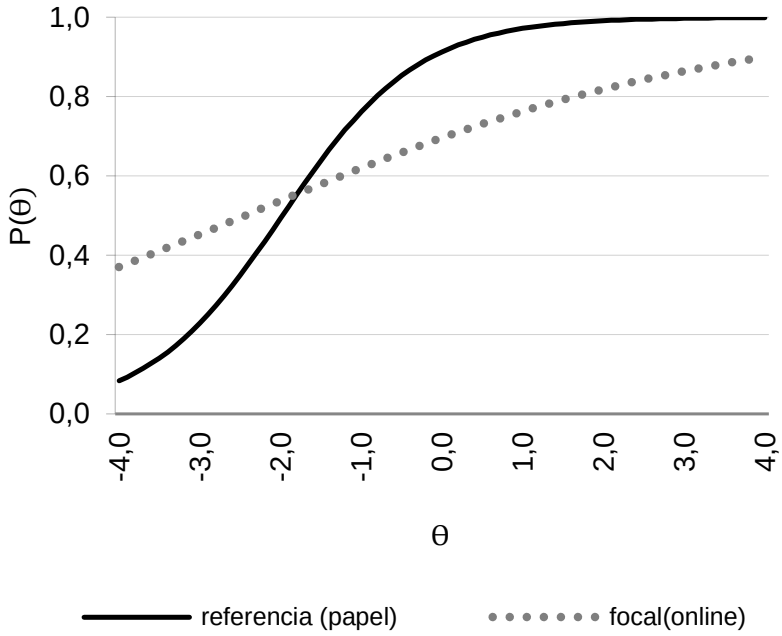


Figura 5. Representación gráfica (CCI) de DIF no uniforme (mixto)

3.1.1.2. Funcionamiento Diferencial de Versiones (DVF)

Cuando se llevan a cabo estudios utilizando la metodología del Funcionamiento Diferencial de los Ítems (DIF) las variables objeto de estudio son de carácter principalmente demográfico, tales como el género (Escorial y Navas, 2006; Gómez e Hidalgo, 1997; Gómez y Navas, 1998; Hamilton, 1999), el tipo de escuela de los estudiantes (Scheuneman y Grima, 1997), etc.

Algunos estudios también trabajan con aspectos relacionados con las características de la prueba; un ejemplo de ello es el procedimiento *Differential Alternative Functioning* (DAF), traducido como *Funcionamiento Diferencial de las Alternativas,* donde se estudia el comportamiento de sujetos en función de algunas variables como género, raza, etnia y conocimiento del constructo latente (Dorans y Holland, 1993; O'Neill y McPeek, 1993; Thissen y Steinberg, 1997).

En cualquier caso, el *Funcionamiento Diferencial de las Alternativas* pretende, al igual que el Funcionamiento diferencial de los Ítems, "*comprender las causas de la*

elección diferenciada de las alternativas de un ítem por los individuos que componen distintos grupos demográficos" (Bandeira, 2002, p.128).

Es en este punto donde surge la propuesta metodológica del presente trabajo, el *"Funcionamiento Diferencial de Versiones"*, un estudio que pretende comparar y verificar si el modo de aplicación de una prueba (versión papel y lápiz vs versión informatizada[7]) produce diferencias en el rendimiento de los estudiantes.

El modo de aplicación no es una variable frecuente en los estudios DIF, puesto que no es de carácter demográfico, pero si se refiere a variables relativas al propio test. Cuando se realiza DIF utilizando como variable principal el modo de aplicación, lo que comúnmente sucede es que un mismo sujeto realiza una prueba en papel y otra prueba online (ambas equivalentes), ya que este procedimiento permite verificar realmente el funcionamiento diferencial de los ítems. Un ejemplo de ello es el estudio llevado a cabo por Poggio et al. (2005), donde se encuentran pocas preguntas de matemáticas (9 de 204 ítems) que se comporten de manera diferente en las dos modalidades.

La metodología DIF requiere de dos grupos de sujetos equivalentes y un ítem que va a ser objeto de estudio. En el ejemplo presentado en la tabla 3.1 llevamos a cabo el estudio del DIF en el ítem 1 para un grupo de sujetos según el género. En dicho ejemplo se dispone de dos puntuaciones: la puntuación en el ítem 1, obtenida por el Grupo de Referencia, es decir por los hombres ($X_{1\ Hombre}$), y la puntuación en el ítem 1, obtenida por el Grupo Focal, las mujeres ($X_{1\ Mujer}$). Con estos datos ya es posible realizar el estudio de la existencia o no de diferencias entre esas puntuaciones.

[7] Esta versión informatiza fue llevada a cabo a través de Internet, por lo que también podemos considerar esta versión como online. De ahora en adelante se utilizará el término online.

Tabla 3.1.
Ejemplo Funcionamiento Diferencial de los Ítems

Ítem	Grupos de sujetos atendiendo al género	
	Grupo de Referencia (GR: Hombre)	Grupo Focal (GF: Mujer)
Ítem 1	$X_{1\ Hombre}$	$X_{1\ Mujer}$

Fuente: Elaboración propia

Dadas las características de nuestro estudio, esta posibilidad es inviable. Disponemos de un grupo de sujetos que realiza exclusivamente la prueba en papel y otro grupo de sujetos que lo hacen exclusivamente online. Esto, unido al inconveniente de que la selección del modo de aplicación no ha sido aleatorio, ha dado como resultado, dos muestras diferentes.

Es necesario que estos grupos sean equivalentes para disponer de dos sujetos con idénticas habilidaes, de forma que uno de ellos realice la prueba en papel mientras que el otro la haga online; sólo así es posible comparar los resultados y que las diferencias observadas se deban exclusivamente al modo de aplicación de la prueba y no a las características del sujeto. Esta idea se aproxima a los estudios del Funcionamiento Diferencial de los Sujetos (DSF), donde se obtienen puntuaciones de cada sujeto en diferentes ítems. Un ejemplo del estudio del DSF puede verse en la tabla 3.2.

Tabla 3.2.
Ejemplo Funcionamiento Diferencial de los Sujetos

Sujetos	Modo de aplicación del Ítem 1	
	Grupo de Referencia (GR: Papel)	Grupo Focal (GR: Online)
Sujeto 1	$X_{1\ Papel}$	$X_{1\ Online}$

Fuente: Elaboración propia

En estos estudios se invierte el proceso y se estudian las diferencias en los sujetos debidas a las características de los grupos que se comparan (Grupo de Referencia–modo aplicación papel; Grupo Focal–modo aplicación online). Se aprecia cómo un mismo sujeto obtiene una puntuación en el ítem en papel y otra puntuación en el ítem online, lo que hace que podamos estudiar el funcionamiento diferencial del sujeto.

Son pocos los estudios DIF que utilizan la idea DSF. Karkee, Kim y Fatiga (2010), y Keng et al., (2008) realizaron un estudio del DIF emparejando a los sujetos en función de pruebas realizadas con anterioridad para poder alcanzar la equivalencia entre los mismos. En este estudio, este procedimiento de nuevo no es posible, pues no disponemos de datos anteriores que nos permitan emparejar a los sujetos en función de sus habilidades.

Por todo lo expuesto en líneas anteriores, podemos considerar la imposibilidad de trabajar de manera estricta con la metodología basada en el DIF tal y como la conocemos, dado que ésta se basa en el estudio de las diferencias en el ítem debidas a las características de los grupos que se comparan (grupo de referencia y focal); y tampoco podemos hacerlo con la metodología basada en el DSF, puesto que no tenemos sujetos que hayan contestado a los mismos ítems en ambas versiones.

La situación con la que nos encontramos (ver tabla 3.3), se caracteriza por disponer de dos puntuaciones. Por un lado, la puntuación obtenida por el sujeto *a* que realiza el ítem 1 en papel ($X_{a\ \text{Papel}}$) y por otro lado, la puntuación obtenida por el sujeto *b* que realiza el ítem 1 online ($X_{b\ \text{Online}}$).

Tabla 3.3.
Ejemplo de la situación de este estudio

Sujetos	Modo de aplicación del Ítem 1	
	Grupo de Referencia (GR: Papel)	Grupo Focal (GR: Online)
Sujeto *a* papel	$X_{a\ \text{Papel}}$	-
Sujeto *b* online	-	$X_{b\ \text{Online}}$

Fuente: Elaboración propia

Ante esta situación, tan solo se podría estudiar la existencia o no de diferencias entre esas puntuaciones. Es por ello que proponemos la siguiente aportación metodológica, que toma por nombre **"Funcionamiento Diferencial de Versiones" (DVF)**, basada en la aproximación a la metodología del Funcionamiento Diferencial de los Sujetos, utilizando el **"Puntaje de Propensión"** o **"Propensity Score"** para lograr disponer de sujetos equivalentes con puntuaciones tanto en la prueba en papel como en la prueba online.

3.1.1.3. Puntaje de Propensión (Propensity Score)

Las características de nuestra muestra hacen que busquemos otros procedimientos para lograr una correcta comparación entre los grupos. Una de las principales complicaciones con la que nos encontramos es la existencia de sesgo de selección, debido a la no selección aleatoria de la muestra: cada centro seleccionaba la versión (papel u online) en que se realizaría la prueba. El resultado fue un elevado número de sujetos que realizaron la prueba en papel (Primaria=9.258; Secundaria=46.482) frente a los que realizan la prueba online (Primaria=1.079; Secundaria=2.207).

Con el fin de superar estas limitaciones, Rosembaum y Rubin (1983), aplicando el modelo causal de Rubin (2005), proponen un método para controlar este sesgo y lograr grupos equivalentes de sujetos comparables. El método es denominado "Puntaje de Propensión", comúnmente conocido en inglés como "*Propensity Score*".

En trabajos recientes como los llevados a cabo por Puhan, Boughton y Kim (2007) y Seo (2013) se realizan estudios comparativos entre dos versiones de un test de rendimiento (papel y online), utilizando para la comparación entre grupos las técnicas de Propensity Score. Los resultados alcanzados demuestran la posibilidad de hablar de equivalencia entre ambas versiones y la no existencia de diferencias estadísticamente significativas en el rendimiento medio de los estudiantes en función del modo de aplicación.

El procedimiento Propensity Score genera una muestra de sujetos que reciben un tratamiento y que es comparable a una muestra de sujetos que no reciben el tratamiento, conocida como grupo de control. Siempre condicionados a unas covariables que nos permitirán estudiar si el efecto se debe a las diferencias reales, propias del modo de aplicación de la prueba, o puedan deberse a otras características ajenas a este suceso.

Para llevar a cabo esta selección y emparejamiento, se requiere de los siguientes pasos (García, 2009):

Paso 1. Estimación y asignación de las Propensity Score a cada individuo.

Antes de proceder a la estimación de las Propensity Score, es necesario disponer de datos adecuados y comparables.

Lo primero que debemos hacer es definir el grupo de control y el grupo tratado:

- Grupo de tratamiento o grupo tratado: son los sujetos que reciben el tratamiento. En nuestro caso se trata de los sujetos que realizan la prueba por ordenador
- Grupo de control: son los sujetos que no reciben el tratamiento. En nuestro caso se trata de los sujetos que realizan la prueba en papel.

A continuación, se procede a la selección adecuada de las covariables que se van a incluir en el modelo. En lo que atañe a ellas, para este estudio se han utilizado las variables *tipo de centro* (público, privado, concertado) y *distrito del área territorial* (Centro, Norte, Sur, Este, Oeste), que son relevantes e influyentes en el ámbito educativo.

Una vez identificados los datos idóneos y seleccionadas las covariables objeto de estudio, se procede a la estimación de las Propensity Score. Matemáticamente, podemos decir que el método Propensity Score se basa en el cálculo de la probabilidad individual de recibir el tratamiento condicionado por ciertas covariables. O lo que es lo mismo, la probabilidad de cada individuo de ser asignado al grupo de tratamiento condicionado por las covariables estudiadas.

La puntuación Propensity Score vendría dada por la siguiente ecuación (Lottridge, Nicewander y Mitzel, 2011):

$$e_i = P\ (T_i = 1\ |X_i) \qquad (2.1)$$

Donde la puntuación PS es:

P = Probabilidad

i = sujeto

$X_i = covariable$

T_i = es la variable Tratamiento, tomará valor 1 cuando el sujeto pertenezca al grupo del tratamiento, y valor 0 cuando no.

Atendiendo a nuestros datos:

$$e_1 = P\ (T_1 = 1\ |X_{TIPOCENTRO}, X_{DAT})$$

P = Probabilidad

1 = sujeto

$X_{TIPOCENTRO}$= covariable Tipo de Centro (público, privado, concertado)

X_{DAT}= covariable Distrito del área territorial (Centro, Norte, Sur, Este, Oeste)

T_i = 1: variable Tratamiento, es decir modo de aplicación Online que se le asignará valor 1 y la aplicación en papel que tendrá valores 0.

Paso 2. Método de emparejamiento

El método de emparejamiento más innovador es el "Emparejamiento por Puntajes de Propensión" o "Propensity Score Matching", que *"tiene el propósito de fortalecer los argumentos sobre la causalidad de las relaciones entre variables"* (Ovalle, 2015, p.82), dado que destacan por resolver problemas derivados de la dimensionalidad y del sesgo de

selección. Por medio del Matching se resume en una sola variable la información ofrecida por un conjunto amplio de variables.

La situación de partida no nos permite conocer la influencia que tiene el tratamiento en el grupo de control y tampoco poder probar la causalidad. Se conoce el resultado del sujeto que realiza la prueba online, pero no es posible conocer cuál es su puntuación en la prueba en papel. El Propensity Score Matching permite evaluar el resultado del tratamiento sin que este se haya aplicado, así como conocer las diferencias que produce el modo de aplicación en el rendimiento de los sujetos.

Las técnicas Matching son métodos ex post-facto, donde emparejamos sujetos con valores similares en las covariables (*tipo de centro* y *distrito del área territorial*) que realicen la prueba online (*Grupo Tratamiento*) y sujetos que lo hagan en papel (*Grupo Control*). Así, encontraremos un sujeto que realice la prueba en papel que sea "similar" a un sujeto que realice la prueba online (Coma, 2012); logrando un balance que nos permitirá comparar grupos homogéneos y ver exclusivamente el efecto que tiene el modo de aplicación de la prueba en el rendimiento.

Las técnicas Matching más comunes y que han sido utilizadas en el estudio empírico de esta tesis son[8]:

Vecino más cercano (K-1)

En esta técnica el procedimiento para llevar a cabo el agrupamiento se caracteriza por asignar a cada individuo del grupo de tratamiento (sujeto que realiza la prueba online) los *k* individuos del grupo de control (sujetos que realizan la prueba en papel) con los valores más próximos en Propensity Score (Westlund, 2013).

[8] Existen otros procedimientos como el Full Matching y Optical Matching que minimizan la distancia entre los sujetos emparejados. Dichos métodos no han sido incluidos en los análisis, pues en la prueba de Secundaria no se consiguió la convergencia del modelo.

En la figura 6, podemos observar un ejemplo de Matching con la técnica de Vecino más cercano 1:2, donde por cada individuo del tratamiento (sujeto que realiza la prueba online) se seleccionan 2 sujetos del grupo control (sujetos que realizan la prueba en papel) con los valores más próximos de Propensity Score.

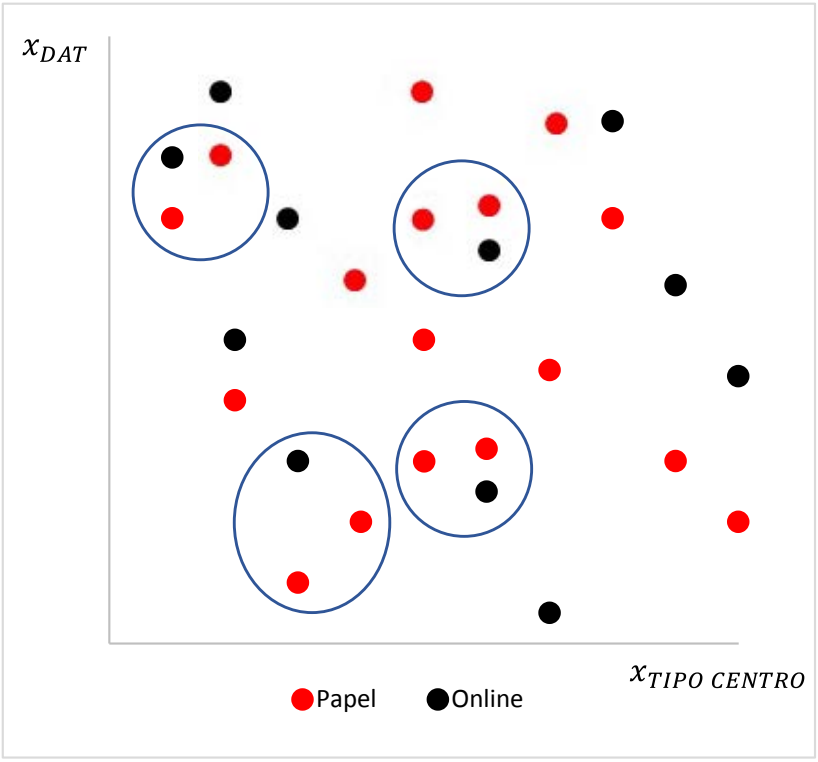


Figura 6. Representación de la técnica Matching del Vecino más cercano con K=2
Fuente: Elaboración propia

Genético o coincidencia genética

Esta técnica atiende a las características del Vecino más cercano y corresponde con el procedimiento que suele usarse, K=1, conocido como Genético o coincidencia genética. Por cada individuo tratado (online) seleccionamos el o los sujetos del grupo control (online) con la misma Propensity Score.

Este procedimiento, por medio de la automatización y de algoritmos de búsqueda genética, encuentra pesos para cada covariable, logrando el mejor equilibrio tras el emparejamiento (Diamond y Sekhon 2013).

El emparejamiento se realiza con la sustitución, por medio de un método de correspondencia llevado a cabo por Abadie e Imbens (2012) basado en pruebas t para variables dicotómicas y una prueba de Kolmogorov-Smirnov para las variables multinomiales y continuas (Westlund, 2013).

Un ejemplo de ello podemos observarlo en la figura 7:

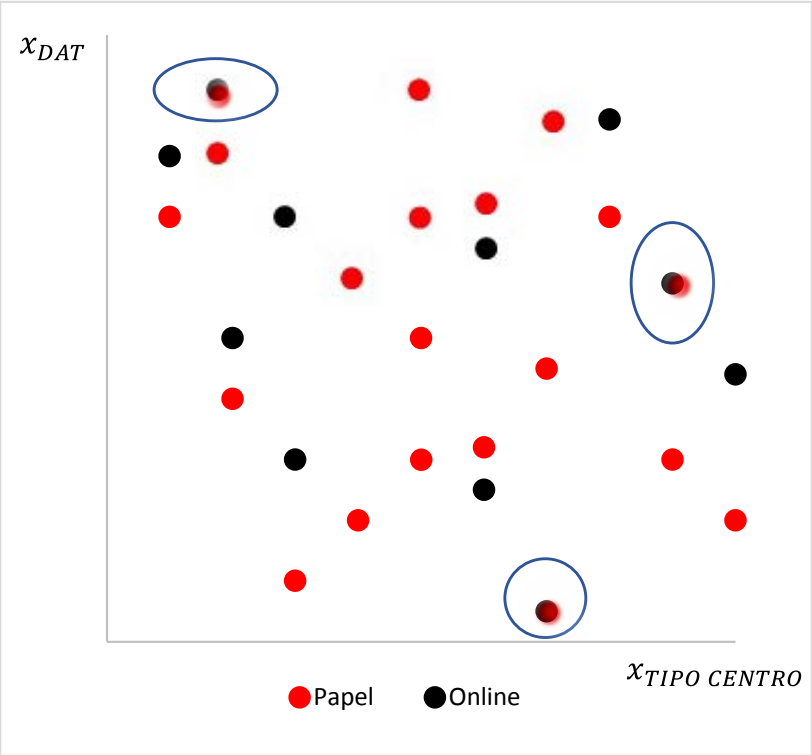


Figura 7. Representación de la técnica Matching de Coincidencia Genética
Fuente: Elaboración propia

Estratificación

Técnica de emparejamiento que agrupa a los sujetos atendiendo a las covariables. Consiste en la creación de subconjuntos en cada uno de los grupos (tratamiento y control), homogéneos entre sí, que posibilitan la comparación de los sujetos. Estos subgrupos homogéneos son conocidos como estratos.

Los sujetos que se encuentran dentro del mismo estrato tienen la misma probabilidad de recibir el tratamiento. Normalmente se utilizan 5 sub-categorías – quintiles (Austin, 2011).

Un ejemplo puede verse en la figura 8, donde los sujetos han sigo agrupados en cinto estratos con similares valores en PS, dejándose fuera los sujetos que no encajan en los grupos.

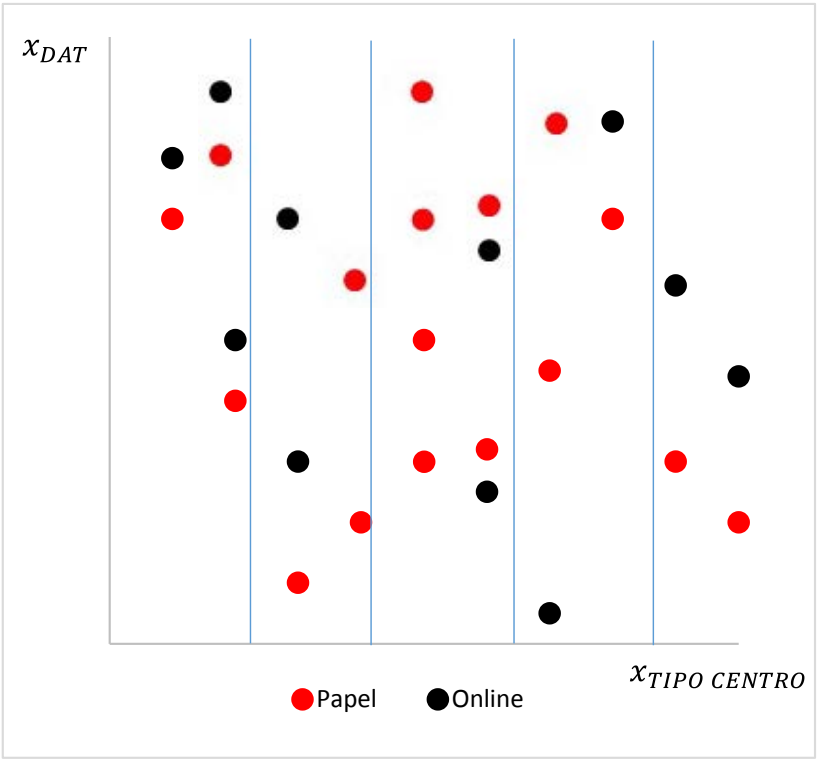


Figura 8. Representación de la técnica Matching de estratificación con 5 etsratos
Fuente: Elaboración propia

Paso 3. Selección del método de emparejamiento

Es recomendable la utilización de varios procedimientos Matching, y analizar el ajuste del balance. Tras llevar a cabo esta comparación, se seleccionará el procedimiento que garantice la mayor similitud entre los sujetos.

La alternativa por la que se ha optado, dadas las características de nuestro estudio, ha sido el emparejamiento de los sujetos que realizan la prueba en papel (grupo de

control), con similares características (mismo distrito y titularidad en el centro), a los sujetos que realizan la prueba en papel (grupo tratado) y que no han podido realizar la prueba online. Por tanto, la transformación que se pretende realizar en este trabajo aplicando la metodología Propensity Score se representa en la siguiente tabla (3.4); consiste en disponer de sujetos equivalentes, y por ende sujetos con puntuación tanto en la versión en papel como en la versión online, para llevar a cabo el estudio del "Funcionamiento Diferencial de Versiones".

Tabla 3.4.
Ejemplo del Funcionamiento Diferencial de Versiones

Sin Propensity Score

Sujetos	Modo de aplicación del Ítem 1	
	Grupo de Referencia (GR: Papel)	Grupo Focal (GR: Online)
Sujeto 1 papel	$X_{1\ Papel}$	-
Sujeto 1 online	-	$X_{1\ Online}$

Con Propensity Score

Sujetos	Modo de aplicación del Ítem 1	
	Grupo de Referencia (GR: Papel)	Grupo Focal (GR: Online)
Sujeto 1	$X_{1\ Papel}$	$X_{1\ Online}$

Nota: Utilizando Propensity Score, obtenemos un único sujeto con puntuación en la prueba en papel y online.
Fuente: Elaboración propia

CAPÍTULO 4: Regresión Logística. Método para la detección y estudio del Funcionamiento Diferencial de Versiones

"El objetivo principal de la educación es crear personas capaces de hacer cosas nuevas, y no simplemente repetir lo que otras generaciones hicieron"

Jean Piaget (1896-1980)

4.1. Introducción

Son muchas las clasificaciones que hay sobre la detección del DIF. A lo largo del tiempo se han ido desarrollando diferentes procedimientos, cada vez más potentes, con la intención de evitar el Error del tipo I (detección errónea de ítems con DIF) y lograr detectar con precisión y exactitud el DIF.

Por este motivo, existen diferentes clasificaciones en función de los métodos de detección del DIF. En la tabla 4.1 se recogen, a modo de resumen, algunas de las clasificaciones propuestas por autores que han abordado el tema, como Bandeira, (2002); Camilli y Shepard (1994); Fidalgo (1996); Gómez e Hidalgo (1997); Herrera (2005); Hidalgo, López-Pina y Sánchez (1997); Mellenbergh (1989); Millsap y Everson (1993); Potenza y Dorans (1995); Santana, (2009); Shepard, Camilli y Averill (1981); Van der Flier, Mellenbergh, Adèr y Wijn (1984); Whitmore y Shumacker (1999).

Tabla 4.1.
Clasificación de los métodos para la detección del DIF

Autores	**Métodos y procedimientos**
Mellenbergh (1989) Van der Flier, Mellenbergh, Adèr y Wijn (1984)	Métodos condicionales: se asume el supuesto de que los parámetros del ítem son diferentes en los grupos de sujetos con la misma habilidad en la variable latente. Métodos incondicionales: se asume el supuesto de interacción entre el grupo y los parámetros de los ítems.
Shepard, Camilli y Averill (1981) Whitmore y Shumacker (1999)	Métodos que utilizan un criterio interno: atendiendo a la puntuación obtenida en los test. Métodos que utilizan un criterio externo: atendiendo a la puntuación obtenida en otro test.

Tabla 4.1.
Clasificación de los métodos para la detección del DIF (continuación)

Autores	Métodos y procedimientos
Camilli y Shepard (1994)	Métodos basados en la Teoría Clásica de los Test (TCT) y en el Análisis de Varianza (ANOVA). Métodos basados en la Teoría de Respuesta al Ítem (TRI): Medidas de área, χ^2 de Lord Comparación modelos/ parámetros Métodos basados en el análisis de tablas de contingencia: Aplicaciones de χ^2, Mantel – Haenszel, Modelos logit y log-lineales, Regresión logística.
Potenza y Dorans (1995)	Según el tipo de criterio de igualación/equiparación de los grupos: Puntuación observada / variable latente Según la relación entre puntuación y la variable de igualación de los grupos: Paramétrica (Método de Regresión Logística y Métodos TRI) No paramétrica (Método de Mantel – Haenszel, Método Estandarizado y procedimiento SIBTEST).
Fidalgo (1996)	Métodos sin un modelo de medida Métodos basados en la TRI
Millsap y Everson (1993) Gómez e Hidalgo (1997) Hidalgo, López-Pina y Sánchez (1997)	Métodos no condicionales Métodos condicionales: Métodos de invarianza condicional observada: χ^2 de Scheuneman, χ^2, Pearson, Método de Mantel-Haenszel. Método estandarizado y el de la regresión logística Métodos de invarianza condicional no observada: utilizan puntuaciones desde el enfoque de la TRI: χ^2 de Lord, el de las áreas Comparación parámetros de los ítems Procedimiento SIBTEST

Tabla 4.1.
Clasificación de los métodos para la detección del DIF (continuación)

Autores	Métodos y procedimientos
Bandeira, (2002)	Basados en la unidimensionalidad de los ítems Método del Delta Gráfico Método del cálculo del área entre las CCI's Método de las Probabilidades Método de la comparación de los parámetros de los ítems Método del Chi-cuadrado de Lord Método del Chi-cuadrado de Sheuneman Método del Chi-cuadrado de Pearson o Total Método de la regresión logística Método de Mantel-Haenszel Método Estandarizado Método Logístico Iterativo Método del análisis de las estructuras de covarianza (MAEC) Basados en la multidimensionalidad de los ítems Método Multidimensional de DIF (MMD) Método de la regresión logística
Herrera (2005)	Procedimientos pioneros: Métodos basados en la Teoría Clásica de los Test (TCT) y en el Análisis de Varianza (ANOVA). Métodos basados en tablas de contingencia Pruebas de hipótesis sobre igualdad de proporciones mediante análisis de tablas de contingencia bidimensional: Aplicaciones χ^2 Método de estandarización Método de Mantel-Haenszel Modelos para el análisis de variables categóricas con tablas de más de dos dimensiones: Modelos log-lineales y logit Método de la regresión logística
Acar y Kelecioglu (2010).	Modelos Jerárquicos Lineales Generalizados (DIF Determining with Hierarchical Generalized Linear Model-HGLM). Regresión Logística. TRI - Likelihood Ratio
Holland y Wainer (2012).	Modelos descriptivos: Método de Mantel-Haenszel y Método de estandarización. Métodos basados en la Teoría de Respuesta al Ítem (TRI)

Fuente: Elaboración propia

4.2. Métodos para la detección del Funcionamiento Diferencial de los Ítems

Considerando la clasificación de Camilli y Shepard (1994), podemos enmarcar los métodos para la detección del funcionamiento diferencial de los ítems en tres grupos: Métodos basados en la Teoría Clásica de los Test (TCT) y en el Análisis de Varianza (ANOVA), métodos basados en la Teoría de Respuesta al Ítem (TRI) y métodos basados en el análisis de tablas de contingencia.

4.2.1. Métodos basados en la Teoría Clásica de los Test (TCT) y en el Análisis de Varianza (ANOVA)

Estos métodos, los primeros propuestos para la detección del DIF, son poco utilizados en la actualidad, dadas sus limitaciones a consecuencia de los avances en los procedimientos y en la utilización de métodos caracterizados por su meticulosidad para la detección del DIF tanto uniforme como no uniforme. A continuación, en la tabla 4.2 se resumen algunas de las ideas de dichos métodos pioneros.

Tabla 4.2.
Descripción de los Métodos basados en la TCT y ANOVA

	Procedimiento (Autor)	**Descripción**	**Algunas ventajas y desventajas**
Métodos pioneros	Teoría Clásica de los Test (Angoff, 1972)	Delta –Plot. Cálculo del índice de dificultad de cada ítem para cada grupo. Siendo un ítem sesgado cuando es más difícil para un grupo que para otro; es decir cuando es mayor la diferencia entre los índices de dificultad en un grupo.	Ventaja: Sencillez en su estimación. Desventaja: Confunde DIF e impacto. Muy dependiente de la muestra, de las características específicas de la muestra, lo que no nos permite establecer generalizaciones ni inferencias
		Correlación Biserial Puntual. Dicha correlación nos permite valorar el comportamiento de los ítems atendiendo a los índices de discriminación de los ítems en cada grupo, a través de la representación gráfica de la correlación biserial puntual.	Ventaja: Sencillez en su estimación. Desventaja: Muy influenciado por la dificultad del ítem y afecta a la correlación dado que esta correlación se caracteriza por ser una medida de relación entre el ítem y el test.

Tabla 4.2.
Descripción de los Métodos basados en la TCT y ANOVA (continuación)

	Procedimiento (Autor)	Descripción	Algunas ventajas y desventajas
Métodos pioneros	ANOVA (Cleary y Hilton, 1968)	ANOVA de medias repetidas de dos factores representado por el grupo de pertenencia como primer factor y los ítems como segundo factor intragrupo.	Ventajas: Sencillez en su estimación. Desventajas: Anticuada. Tiene falsos positivos y valores altos de errores tipo I y II. El motivo de estos fallos en la detección, es que tan solo atiende a la dificultad de los ítems y el impacto entre los grupos comparados pero se olvida de la discriminación de cada uno de los ítems.

Fuente: Adaptación Cuevas (2013, p.24)

A continuación detallaremos el procedimiento basado en la Teoría Clásica de los Test utilizado en este trabajo para la detección del DVF: Transformación del Índice de Dificultad (T.I.D.).

Delta-plot / Transformación del Índice de Dificultad (T.I.D)

Método propuesto por Angoff en 1972, basado en el cálculo del índice de dificultad de cada ítem para cada grupo, siendo un ítem sesgado cuando es más difícil para un grupo que para otro; es decir, cuando es mayor la diferencia entre los índices de dificultad en un grupo. Por ello, se obtiene la dificultad diferencial o incremental del ítem, denominación que tuvo en un principio el sesgo (Angoff, 1993).

Angoff y Ford (1973, citados por Magis y Facon, 2012) proponen la transformación de los índices de dificultad calculados para cada ítem (p), recibiendo el nombre de índices de dificultad transformados o método del delta gráfico, muy utilizado por su sencillez durante los años 70.

Una vez llevada a cabo esta transformación del índice de dificultad se elabora un diagrama de dispersión que refleje en el eje de ordenadas los valores de *p* transformados para cada ítem del grupo focal, y en el eje de abscisas los valores de

p transformados para cada ítem del grupo de referencia. Este gráfico nos permite conocer los ítems que podrían ser diagnosticados como sesgados, y corresponde a aquellos ítems que se alejan del eje principal, recta donde se encuentran la mayoría de los ítems en forma de una elipse alargada y estrecha (Gómez e Hidalgo, 1997).

Además de esta representación gráfica, los autores proponen un índice de distancia perpendicular de cada ítem al eje principal, la recta (los valores más altos, sean positivos o negativos, tendrán funcionamiento diferencial).

Siendo la ecuación de la recta: y = ax + b:

$$a = \frac{(s_y^2 - s_x^2) \pm \sqrt{(s_y^2 - s_x^2) + 4r_{xy}^2 s_x^2 s_y^2}}{2r_{xy}s_x s_y} \qquad (5.1)$$

$$b = M_y - aM_x$$

Siendo:

x e y: valores de Δ para los grupos comparados
M_y y M_x: medias de los grupos
s_y y s_x: desviaciones típicas de los grupos
r_{xy}: correlación de Pearson de los grupos.
La distancia al eje principal, recta de cada punto i es:

$$d_i \frac{ax_i - y_i + b}{\sqrt{a^2} + 1}$$

Error típico para Δ:

$$\sigma_{\Delta ij} = \frac{4}{N_j - 1}$$

El error típico para Δ es constante ≈ 0,01.
N_j: número de sujetos en el nivel j

Magis y Facon (2012, 2013, 2014) detallan en sus trabajos este procedimiento de detección del DIF, así como sus mejoras y la potencia de dicho procedimiento con muestras pequeñas.

El inconveniente de este método es que tiende a confundir el DIF con el impacto cuando los grupos realmente poseen DIF. Además, se trata de un índice muy dependiente de las características específicas de la muestra, lo que no nos permite establecer generalizaciones ni inferencias (Shepard, Camilli y Averill, 1981, citados por Gómez e Hidalgo, 1997).

4.2.2. Métodos basados en la Teoría de Respuesta al Ítem (TRI)

Estos métodos, como indica Cromwell (2006, p.16, citado en Santana 2009, p.25), consisten en "*determinar si hay diferencia en los parámetros de los ítems entre el grupo focal y el grupo de referencia*", es decir, un ítem presentará DIF cuando las funciones de respuesta al ítem sean diferentes en ambos grupos, el de referencia y el focal.

Los principales métodos basados en la Teoría de Respuesta al Ítem para la detección del DIF, según Herrera (2005), pueden agruparse en tres grupos: el primero basado en la comparación de los modelos ajustados, el segundo un grupo que para la detección del DIF estudia la comparación de los parámetros de los modelos y, por último, los procedimientos centrados en el cálculo de las medidas del área de discrepancia entre las CCIs de ambos grupos.

Un resumen de estos procedimientos lo podemos observar en la tabla 4.3.

Tabla 4.3.
Descripción de los Métodos basados en la TRI

	Procedimiento (Autor)	Descripción	Algunas ventajas y desventajas
Métodos basados en la TRI		Comparación de medidas del área	
	Medidas de área entre las CCI. Rudner (1977). Rudner, Getson y Knight (1980)	Evalúa el área entre las dos CCI de un ítem que son ajustadas para dos grupos de forma independiente y expresadas en la misma métrica (Herrera et al., 2007).	Ventajas: Sencillo y fácil de calcular. Desventajas: Exigencias en el tamaño de muestra. Decisión subjetiva (no existe unanimidad en considerar el índice como alto o bajo), no hay prueba de significación estadística.
	Rajú (1988)	Muchos autores proponen diferentes índices para medir el área entre las CCI. Rajú (1988) propone otro estadístico para el cálculo del DIF, basado en el cálculo del área entre las CCI´s en modelos logísticos de uno, dos y tres parámetros	
		Comparación de parámetros	
	Lord (1980)	Compara los vectores de los parámetros estimados cuando se ajustan modelos TRI para dos grupos de forma separada (Herrera et al., 2007).	Ventajas: Procedimiento sencillo de utilizar, cuenta con una prueba estadística y algunos estudios aplicados respaldan su uso. Desventajas: Dificultades de aplicación con modelos de tres parámetros. Exigencias en tamaños de muestra, supone que θ es conocido y es aplicable únicamente con algoritmos de máxima verosimilitud. Desconocimiento del tamaño de muestra necesario para lograr la convergencia a la distribución y posible alta tasa de falsos positivos (Herrera et al., 2007).
	Diferencia del parámetro b Wrignt el al. (1976)	Comparación de los parámetros de dificultad para dos grupos controlando por el nivel de habilidad.	Ventajas: Sencillo y fácil de calcular. Desventajas: Debilidades para detectar DIF no uniforme y puede ser engañoso en situaciones en las que se ajusten mejor modelos de dos o tres parámetros (Camilli y Shepard, 1994).
	Comparación de modelos (Thissen, Steinberg y Gerrard, 1986)	Comparación de modelos TRI para detectar DIF (Herrera, 2005).	Ventajas: Sencillo y fácil de calcular. Desventajas: Incapacidad para la detección de diferencias pequeñas. Dependencia del tamaño muestral de los grupos (Gómez Benito e Hidalgo Montesinos, 1997).

Fuente: Adaptación Cuevas (2013, p.24)

En las siguientes líneas profundizaremos sólo en aquellos procedimientos utilizados en este trabajo para la detección del DVF. Concretamente, para este estudio se han utilizado como métodos basados en la TRI el método Chi cuadrado de Lord y el índice de Rajú.

Método Chi cuadrado de Lord

En lugar de comparar cada parámetro por separado, Lord (1980, citado en Núñez, Hidalgo y López, 2000) propone que se comparen conjuntamente, atendiendo a los vectores de parámetros estimados una vez que se ha llevado a cabo el ajuste de los modelos (un parámetro, dos o tres) en cada grupo.

La manera de llevarlo a cabo, es por medio del estadístico Chi-cuadrado (χ^2).

$$\chi^2 = V' \sum^{-1} \mathrm{V} \qquad (5.7)$$

χ^2: Chi-cuadrado de Lord (dos grados de libertad)
V': vector transpuesto de V
$\sum^{-1}$: es la inversa de la matriz de varianza-covarianza asintótica para los vectores de diferencias entre parámetros.
V: vector de diferencias entre los parámetros estimados para un ítem en el grupo de referencia y los parámetros estimados para ese mismo ítem en el grupo focal.

Índice de Rajú (métodos basados en el cálculo de las medidas del área)

Este método, basado en la TRI, se diferencia de los anteriores en que para la detección del DIF no se consideran ni los modelos ni los parámetros de los ítems en cada grupo, sino que se tiene en cuenta el área entre las CCIs del grupo de referencia y del grupo focal (Herrera, 2005), para lo que se representan gráficamente las CCIs de ambos. Si ambas curvas no coinciden, los ítems presentan DIF.

Para establecer la diferencia entre las CCIs de ambos grupos es necesario calcular dicha área entre las CCIs del grupo de referencia y las del grupo focal. Rudner (1977), y Rudner, Getson y Knight (1980), citados en Herrera (2005), proponen un índice para calcular el área comprendida entre ambas CCIs, en los valores de (θ) comprendidos en el intervalo (-3 < θ <+3).

$$A_i = \sum_{\theta=-3}^{\theta=+3} |P_R(\theta) - P_F(\theta)| \Delta_\theta \qquad (5.10)$$

Siendo:
$P_R(\theta)$: valor de la probabilidad de acertar el ítem en el grupo de referencia
$P_F(\theta)$: valor de la probabilidad de acertar el ítem en el grupo focal
$P_R(\theta) - P_F(\theta)$: la recta de estas probabilidades da lugar a la altura de un rectángulo
Δ_θ: es el área de un rectángulo
$\theta = -3\ a + 3$: los valores de θ están comprendidos entre -3 y +3.

Si obtenemos un valor alto de A_i el ítem funciona diferencialmente, si $A_i = 0$ el ítem no presenta DIF, pero no sabemos exactamente qué se puede considerar valor alto, por lo que se trata de una decisión un tanto subjetiva.

Por ello, como señala Herrera (2005), para tratar de solventar estas limitaciones son muchos los autores que proponen índices, como los índices de diferencia de probabilidad, la suma de cuadrados autoponderados, la medida de área con signo o los cuatro índices de DIF (área base superior, área base inferior, área total, RDMC), entre otros.

Rajú (1988, citado en Núñez, Hidalgo y López, 2000) propone otro estadístico para el cálculo del DIF, basado en el cálculo del área entre las CCIs en modelos logísticos de uno, dos y tres parámetros:

$$A = (1 - c)\left|\frac{2(a_r a_f)}{D a_r a_f} \ln\left[1 + e^{\frac{D a_r a_f (b_r b_f)}{(a_r a_f)}}\right] - (b_r - b_f)\right| \qquad (5.11)$$

$a_{r:}$ parámetro de dificultad del grupo de referencia
$a_{f:}$ parámetro de dificultad del grupo focal
$b_{r:}$ parámetro de discriminación del grupo de referencia
$b_{f:}$ parámetro de discriminación del grupo focal

c: probabilidad de acierto por azar
D: constante de valor 1,7
e: base logaritmos neperiano, 2,7182

Esta fórmula puede transformarse mediante integrales:

$$A = \int [P_R(\theta) - P_F(\theta)]\, d_\theta \quad (5.12)$$

Pero este estadístico no diferencia correctamente el DIF no uniforme, para solucionar este problema, se eleva las diferencias al cuadrado y se extrae la raíz cuadrada de la fórmula anterior:

$$A = \sqrt{\int [P_R(\theta) - P_F(\theta)]^2\, d_\theta} \quad (5.13)$$

Swaminathan y Rogers (1990), simplifican la fórmula anterior, cuando los parámetros c y a son iguales en ambos grupos:

$$A = (1 - c)|\,(b_r - b_f)| \quad (5.14)$$

En modelos de dos parámetros no se atiende al parámetro c.
En modelos de un parámetro, se tiene en consideración únicamente la diferencia entre la dificultad:

$$A = |(b_r - b_f)| \quad (5.15)$$

4.2.3. Métodos basados en el análisis de tablas de contingencia.

Siguiendo con la clasificación llevada a cabo por Camilli y Shepard (1994), pasamos ahora a explicar brevemente algunos de los métodos más comunes basados en el análisis de las tablas de contingencia.

Herrera (2005) hace una clasificación muy acertada que recogemos en las siguientes líneas: clasifica estos métodos según se fundamenten en pruebas de hipótesis sobre igualdad de proporciones mediante análisis de tablas de contingencia bidimensional, y los que originan modelos para el análisis de variables categóricas con tablas de más de dos dimensiones.

En la tabla 4.4 se muestra un resumen de dichos procedimientos.

Tabla 4.4.
Descripción de los Métodos basados en tablas de contingencia

	Procedimiento (Autor)	Descripción	Algunas ventajas y desventajas
Métodos basados en el análisis de tablas de contingencia		<u>Métodos basados en pruebas de igualdad de proporciones</u>	
	Aplicaciones de χ^2 (Scheuneman,1981)	Plantea que se puede descartar la presencia de DIF si la proporción de aciertos es igual tanto el grupo de referencia como el grupo focal en los diferentes estratos o niveles. Es por ello, que este procedimiento también es conocido como el Chi cuadrado de los aciertos (Santana, 2009).	Ventajas: Sencillez y economía. Desventaja: Incapacidad para detectar DIF no uniforme. Solo tiene en cuenta la proporción de aciertos, lo que puede llevarnos a resultados erróneos concretamente si existe impacto y diferencia muestral entre los grupos.
	Método de estandarización. (Kulick y Dorans, 1983)	Estudio de las diferencias entre las proporciones de acierto que se dan entre los dos grupos de comparación, de referencia y focal, cuando están sujetos a un mismo grado de habilidad.	Ventajas: Cuenta con una métrica y prueba de hipótesis. Desventajas: Únicamente utiliza la proporción de aciertos lo que puede distorsionar los resultados.
	Mantel Haesnzel (Holland y Thayer 1986, 1988)	Si un ítem no presenta DIF, la razón entre el número de personas que lo fallan y lo aciertan es igual para los grupos en diferentes niveles de habilidad (Herrera et al., 2005).	Ventajas: Sencillez y economía computacional. Es eficiente al manejar diferentes niveles de habilidad como variable de control. Cuenta con una métrica y prueba estadística. Desventajas: Posible contaminación de los ítems DIF, cuando se utilizan los puntajes observados en la prueba como criterio de pareamiento de los grupo.
		<u>Modelo para el análisis de tablas</u>	
	Modelos log-lineales y modelos logit (Mellenbergh,1982)	Estos modelos buscan ajustar un modelo cuyo fin es predecir la frecuencia esperada de cada celda de la tabla de contingencia como producto de los efectos incluidos en el modelo. (Herrera et al., 2005).	Ventaja; Permite describir las características de las variables que conforman el modelo y sus interacciones. Desventajas: Posible contaminación de los ítems DIF, cuando se utilizan los puntajes observados en la prueba como criterio de pareamiento de los grupo.
	Regresión logística. (Spray y Carlosn, 1986; Bennet, Rock y Kaplan, 1987; Swaminathan y Togers, 1990).	Caso particular del análisis de regresión múltiple cuando la variable dependiente es dicótoma (Herrera et al., 2005).	Ventajas: Facilidad para ajustarse al análisis de ítems politómicos o a situaciones en las que se tienen más grupos. Capacidad para detectar DIF uniforme y DIF no uniforme. Desventajas: Posible contaminación de los ítems DIF, cuando se utilizan los puntajes observados en la prueba como criterio de pareamiento de los grupo.

Fuente: Adaptación Cuevas (2013, p.24)

Dado que el estudio del DVF presentado aborda algunos de estos procedimientos (método Estandarizado-Stand y Mantel–Haenszel), a continuación explicaremos brevemente en qué consisten.

Método de estandarización

Método propuesto por Dorans y Kulick en 1986, también conocido como Diferencia de Proporciones Estandarizadas (D_{EST}).

Este procedimiento consiste en el estudio de las diferencias entre las proporciones de acierto que se dan entre los dos grupos de comparación, de referencia y focal, cuando están sujetos a un mismo grado de habilidad. Estas diferencias son estudiadas dividiendo el continuo en K intervalos y atendiendo a las diferencias en las proporciones de aciertos que tiene cada grupo en cada puntuación (j) y en cada intervalo (k) en los que se ha dividido el rango total (Gómez e Hidalgo, 1997).

Dorans y Holland (1993) plantean el siguiente estadístico:

$$D_{EST} = \sum_{j=1}^{K} w_j \, \Delta p_j / \sum_{j=1}^{K} w_j \qquad (5.18)$$

Siendo:
Δpj= $p_{fj} - p_{rj}$: La diferencia entre la proporción de aciertos de ambos grupos
w_j: factor de ponderación para los grupos estudiados.
p_{fj} y p_{rj} : proporción de sujetos que aciertan el ítem comparadas con los que lo aciertan en el grupo focal y de referencia.

Toma valores entre -1 y +1 (favoreciendo al grupo de referencia cuando los valores son negativos y favoreciendo al grupo focal cuando los valores son positivos).

El ítem no presenta DIF o es irrelevante cuando D_{EST} se encuentra entre los valores -0,05 y + 0,05
El ítem puede presentar DIF moderado cuando D_{EST} se encuentra entre los valores -0,10 y -0,05 o entre +0,10 y +0,05.
El ítem presenta DIF severo cuando D_{EST} se encuentra entre los valores inferiores a -0,10 y superiores a +0,10.

Método de Mantel Hanszel

Procedimiento elaborado en 1959 por N. Mantel y W. Haenszel, aunque no fue llevado a la práctica en psicometría hasta 1988 por P.W. Holland y D. T. Thayer (Bandeira, 2002, 2003).

Este procedimiento consiste en la detección del Funcionamiento Diferencial de los Ítems, comparando las frecuencias observadas y esperadas de aciertos y errores en el Grupo de Referencia (alumnos que hacen la prueba en papel) y Grupo Focal (alumnos que realizan la prueba online), de acuerdo con los distintos niveles de habilidad elegidos por el investigador (j) en la variable latente (θ) (Bandeira, 2002). Por tanto, un ítem no presenta DIF cuando es igual la razón entre los sujetos que aciertan y fallan tanto en el grupo de referencia como en el grupo focal para los estratos fijados (Herrera, 2005).

Esta razón, representada como αMH, es conocida con el nombre de *"Odds ratio"* (Herrera, 2005): expresa la razón entre la probabilidad de acertar el ítem *versus* la de fallarlo en el grupo de referencia y en el grupo focal (Hidalgo, López y Sánchez, 1997).

El estadístico αMH, también nombrado Chi Cuadrado de Mantel-Haenszel, sigue una distribución χ^2 con un grado de libertad, y representará ausencia de DIF cuanto obtenga un valor igual a 1 (H_0: αMH =1), aunque, por el contrario, determinará presencia de DIF cuando obtenga un valor diferente a 1 (H_1: αMH ≠1). Concretamente, si es >1 el ítem favorece al grupo de referencia, y si <1 el ítem favorece al grupo focal.

Formulación matemática:

$$\alpha MH = \frac{\frac{\sum_{j=1}^{S} A_j D_j}{T_j}}{\frac{\sum_{j=1}^{S} B_j C_j}{T_j}} \qquad (5.19)$$

Siendo:

A_j : la frecuencia observada de las respuestas correctas del grupo de referencia en el nivel j de la puntuación observada.

D_j : la frecuencia observada de las respuestas correctas del grupo focal en el nivel j de la puntuación observada;
B_j : la frecuencia observada de las respuestas incorrectas del grupo de referencia en el nivel j de la puntuación observada;
C_j : la frecuencia observada de las respuestas incorrectas del grupo focal en el nivel j de la puntuación observada;
T_j : número de sujetos en el nivel j.

Para facilitar la interpretación de este coeficiente (dado que toma valores entre 0 y ∞) Holland y Thayer (1988) propusieron la transformación logarítmica del coeficiente αMH, comúnmente conocida como *"escala delta"* (Δ_{MH}), dando lugar a la ausencia de DIF cuando Ln (αMH) = 0. Si, por el contrario, Ln (αMH) ≠ 0, entonces se puede hablar de la presencia de DIF. Más concretamente: si Ln (αMH) < 0 el ítem favorece al grupo de referencia, mientras que si Ln (αMH) > 0 el ítem será más fácil para el grupo focal.

Para una interpretación correcta, es importante hablar de una escala jerárquica para los valores del coeficiente Ln (αMH), como la propuesta por el Educational Testing Service, que nos permite interpretar el tamaño del DIF de la siguiente manera (Cuevas, 2013; Dorans y Holland, 1993; Elosua, 2006; Longford, Holland y Thayer, 1993; Roussos, Schnipke y Pashley, 1999; Ziecky, 1993):

$|\text{Ln}(\alpha MH)| = 0$ ó < 1 (DIF irrelevante)
$|\text{Ln}(\alpha MH)| \leq 1{,}0$ o $1{,}5 > |\text{Ln}(\alpha MH)| > 1$ (DIF moderado)
$|\text{Ln}(\alpha MH)| > 1{,}5$ (DIF severo)

A continuación abordaremos con mayor detalle el método de regresión logística, utilizado en este estudio para la detección del DVF.

4.3. Regresión Logística para detectar DVF

Se trata de un método para la detección del DIF o, en nuestro caso, para detectar DVF, basado en el análisis de tablas de contingencia.

Herrera (2005) clasifica esta técnica dentro de los métodos basados en el análisis de tablas de contingencia, mientras que Swaminathan y Rogers (1990) sugieren la utilización de regresión logística como alternativa a los métodos de la TRI. Es éste un método superior a otros, puesto que estudia tanto el DVF uniforme como el DVF no uniforme, además de atender a la naturaleza continua de la escala de habilidad (Gómez y Hidalgo, 1997; Hidalgo y López-Pina, 2004; Zumbo, 1999).

Este procedimiento, en palabras de Santana (2009), *"considera la probabilidad de acierto a un ítem como función de la habilidad del sujeto, y el grupo al cual pertenece el sujeto […] evalúa la interacción entre la habilidad y el grupo de pertenencia, y su influencia en la probabilidad de acierto al ítem"* (p.29). Se valora la presencia de DVF a través del estudio de la mejora en el ajuste que produce la incorporación sucesiva al modelo de regresión logística.

Matemáticamente, podemos recoger estas ideas en la siguiente expresión (Swaminathan y Rogers, 1990):

Formulación matemáticamente:

$$p(y = 1|\theta) = \frac{e^z}{1+e^z} \qquad (3.1)$$

Dónde:

p (y = 1| θ): probabilidad de acertar el ítem dado un nivel de atributo θ,
Z: combinación lineal de las variables predictoras de esa probabilidad de acierto que puede expresarse matemáticamente como:

$$Z = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g \qquad (3.2)$$

Dónde:

θ= nivel de habilidad del sujeto en la prueba (puntuación en la prueba)
g= grupo al que pertenece el sujeto (grupo de referencia y grupo focal)
θ g= interacción entre el nivel de habilidad y el grupo
β_0= intercepto
$\beta_1, \beta_2, \beta_3$: Coeficientes para la habilidad, el grupo y la interacción grupo – habilidad, respectivamente.

Interpretación:

$\beta_2 \neq 0$ y $\beta_3 = 0$ (ítem presenta DVF uniforme)
$\beta_3 \neq 0$ (Ítem presenta DVF no uniforme)

Para la estimación de los parámetros de cada uno de los modelos se utiliza el método de máxima verosimilitud, logrando valores en los parámetros de cada uno de los modelos que maximicen la función de verosimilitud [-2Log] (Alderete, 2006; Thomas y Zumbo, 1996).

Cuando estamos ante modelos con variables dicotómicas nos enfrentamos a serios problemas relacionados con el cumplimiento de los supuestos de regresión lineal, dado que *"la respuesta dada por la probabilidad de un evento no es lineal"* (Santana, 2009, p.31). Esto, unido a la dificultad de calcular los parámetros, supuso el empleo de la transformación logit presentada a continuación.

Formulación matemática:

$$logit(p) = ln\left[\frac{p}{1-p}\right] \quad (3.3)$$

Dónde:

p: probabilidad de acierto de un suceso
1-p: probabilidad de fracaso de un suceso

Los parámetros que debemos estimar en cada uno de los modelos son los siguientes:

Modelo 1: Habilidad o Puntuación Total.

- Únicamente valora el parámetro de la variable habilidad o puntuación total.

Modelo 2: Habilidad o Puntuación total + Grupo.

- Incluye al modelo 1, el parámetro de pertenencia al grupo. Se estudia el DVF uniforme, cuando la diferencia entre el modelo 1 y el modelo 2 es significativa.

Modelo 3: Habilidad o Puntuación total + Grupo + Interacción entre habilidad o Puntuación total y Grupo.

- Incluye al modelo 2, el término de interacción entre la habilidad o puntuación total y el grupo. Cuando se compara el modelo 2 y 3 se estudia el DVF no uniforme

Para el estudio del DVF podemos llevar a cabo varios análisis. Uno de ellos, conocido como *"comparación de modelos anidados"*, consiste en que, una vez se dispone de los modelos, se procede al estudio de la significación estadística para conocer la bondad de ajuste (razón de verosimilitud) de las variables incorporadas en cada uno de los modelos sometidos al test $\chi 2$ (gl = 2) (Santana, 2009).

Cuando la prueba $\chi 2$ para la diferencia en los modelos 1 y 2 muestra diferencias significativas ($p<0,01$) nos indica que no hay DVF uniforme; en cambio, cuando no muestra diferencias significativas ($p>0,01$) nos indica que hay DIF uniforme.

Para el estudio del DVF no uniforme se utiliza la prueba $\chi 2$ para la diferencia en los modelos 2 y 3. Si dichas diferencias son significativas ($p<0,01$) nos indica que no hay DVF no uniforme, pero si no muestra diferencias significativas ($p>0,01$) indica que hay DVF no uniforme.

Existe otro análisis, conocido como *"prueba simultánea de la presencia de DVF uniforme y no uniforme"*. Este procedimiento es el más utilizado y el que obtiene mejores resultados en la detección correcta del DVF y en el control del error tipo I (Herrera, 2005; Santana, 2009; Swaminathan y Rogers, 1990).

Cuando la prueba $\chi 2$ para la diferencia en los modelos 1 y 3 muestra diferencias significativas ($p<0,01$) nos indica presencia de DVF. Si no muestra diferencias significativas ($p>0,01$) representa la ausencia de DVF.

Para considerar la magnitud de esta diferencia, Zumbo y Thomas (1996) propusieron el uso de una medida del efecto del DVF basada en la medida de mínimos cuadrados ponderados (Pseudo – R^2).

Los coeficientes de regresión β_2 y β_3 pueden ser considerados como medidas de magnitud del DIF uniforme y no uniforme, respectivamente (Choi, Gibbons y Crane, 2011 y Jodoin y Gierl, 2001); mientras que la diferencia en el coeficiente β_1 de los modelos 1 y 2 también se ha considerado como medida de magnitud del DIF uniforme (Crane, Belle y Larson 2004).

Por otro lado, la medida del efecto del DVF (R^2) *"representa la proporción de variación de las respuestas al ítem explicada por el modelo de regresión"* (Elosua y López-Jauregui, 2007, p.331), y nos permite conocer la intensidad del DVF. Según la literatura, y siguiendo a estos mismos autores, podemos considerar los siguientes valores para interpretar la medida del efecto del DVF:

DVF débil o insignificante cuando los valores de $R^2 < 0,035$;
DVF moderado: $0,035 < R^2 < 0,07$ y
DVF relevante o elevado: $R^2 > 0,07$.

Para llevar a cabo las comparaciones entre grupos y la detección del DVF es imprescindible realizar estas comparaciones entre sujetos con el mismo nivel de rasgo

latente. Existen procedimientos que establecen estos niveles a través de la puntuación total en la prueba, pero para garantizar la selección correcta de los sujetos con los mismos niveles de rasgo latente utilizaremos en este estudio las puntuaciones TRI junto a la regresión logística para la detección del DVF.

PARTE 2: TRABAJO EMPÍRICO

CAPÍTULO 5: Presentación del problema y de las hipótesis de investigación

”La verdadera ciencia enseña, por encima de todo, a dudar y a ser ignorante”.

(Miguel de Unamuno)

En los siguientes capítulos, correspondientes a la parte empírica del trabajo, daremos respuesta al problema de investigación y las hipótesis planteadas tras la revisión bibliográfica llevada cabo en el marco teórico presentado. También se mostrarán los resultados y conclusiones derivadas del estudio empírico.

5.1. Delimitación del problema de investigación

Tal y como se ha señalado en el capítulo 2, la International Test Commission (2005), propone algunas directrices que nos han guiado en el planteamiento del problema y de las hipótesis de investigación.

Las *"Directrices de los test informatizados"* (ver anexo 6), estrictamente relacionadas con la equivalencia del test informatizado y el test convencional, nos señalan que:

> *Las puntuaciones provenientes de aplicaciones convencionales y de test informatizados pueden considerarse equivalentes cuando:*
>
> > *a) el rango de las puntuaciones de las personas en ambas formas es muy similar;*
> >
> > *b) las medias, variabilidad y forma de las distribuciones de las puntuaciones son aproximadamente las mismas o pueden hacerse similares mediante un reescalamiento de las puntuaciones de la versión informatizada.*
>
> *Quienes realizan una versión informatizada de un test convencional deben aportar los datos sobre su validez.*
>
> *El constructor de un test debería proporcionar estudios comparativos de las versiones convencional e informatizada para establecer la fiabilidad relativa de la aplicación informatizada (*APA, 1986, 2014; Muñiz y Hambleton, 1999 y Lorenzo, 2003).

Por todo ello, para garantizar la equivalencia entre las versiones, es necesario realizar un estudio de validez y fiabilidad, que debe ofrecer resultados semejantes.

Esto, unido a la evaluación del Funcionamiento Diferencial del Ítem, permitirá detectar posibles contrariedades, identificar dónde se producen, en qué grupos y por lo tanto, modificar el test y superar esos problemas (ITC, 2005; AERA, APA, y NCME, 2014).

Tal y como se ha venido adelantando en el marco teórico, el problema de investigación planteado en esta tesis se centra en los test informatizados y la influencia que puede tener en el rendimiento educativo. Revisadas las investigaciones sobre la equivalencia de las versiones (papel y lápiz y online), nos planteamos los siguientes interrogantes:

> ***"¿El modo en que es aplicado un test (papel y online) provoca diferencias estadísticamente significativas en el rendimiento en comprensión lectora?"***
>
> ***"¿El modo en que es aplicado un test (papel y online) provoca Funcionamiento Diferencial de Versiones en la prueba de rendimiento en comprensión lectora?"***

Antes de dar respuesta a estos problemas de investigación, debemos atender el primer objetivo marcado, en torno al cual gira la tesis doctoral.

Objetivo 1

"Demostrar la utilidad de las Propensity Score para el emparejamiento efectivo de muestras y para el estudio de la equivalencia y de la detección del Funcionamiento Diferencial de Versiones en ambas versiones de una prueba de rendimiento en comprensión lectora".

Una vez desarrollado el primer objetivo propuesto, para dar respuesta científicamente a los problemas de investigación formulados, mostramos el resto de objetivos propuestos y las hipótesis a verificar en el estudio empírico llevado a cabo.

5.1.1. Objetivos e hipótesis asociados a la evaluación de la equivalencia de la prueba de rendimiento en Comprensión lectora en dos versiones (papel y online).

Objetivo 2

"Evaluar la equivalencia de una prueba de rendimiento en comprensión lectora elaborada en dos versiones (papel y online).

Hipótesis 1

"No existen diferencias en las características psicométricas (atendiendo a la Teoría Clásica de los Test) entre la versión en papel y online de la prueba de rendimiento en comprensión lectora".

Hipótesis 2

"No existen diferencias en los parámetros de la Teoría de Respuesta al Ítem en la versión en papel y online de la prueba de rendimiento en comprensión lectora".

Hipótesis 3

"La estructura factorial es invariante en la versión en papel y online de la prueba de rendimiento en comprensión lectora".

Hipótesis 4

"No existen diferencias estadísticamente significativas en la distribución de las puntuaciones (media y varianza) en la versión en papel y online de la prueba de rendimiento en comprensión lectora, por tanto ambas versiones son equivalentes".

Hipótesis 5

"No existe asociación entre la puntuación en cada ítem y su modo de aplicación, ya que la media de cada ítem en ambas versiones son iguales, por tanto ambas versiones de la prueba de comprensión lectora son equivalentes".

5.1.2. Objetivos e hipótesis asociados a la detección del Funcionamiento Diferencial de Versiones en la prueba de rendimiento.

Objetivo 3

"Evaluar la equivalencia de una prueba de rendimiento en comprensión lectora elaborada en dos versiones (papel y online) por medio del estudio del Funcionamiento Diferencial de Versiones.

Hipótesis 6

"Teniendo en cuenta el modo en el que se ha aplicado el test (papel y online), no existen Funcionamiento Diferencial de Versiones, por tanto ambas versiones de la prueba de rendimiento en comprensión lectora son equivalentes".

CAPÍTULO 6: Diseño de la Investigación

"Aprender es descubrir que algo es posible"

Fritz Perls (1893 – 1970)

El diseño utilizado en este trabajo es una investigación mediante encuesta, técnica cuantitativa que nos permite recoger información sobre variables concretas. Esta recolección de la información ha sido posible gracias al uso de un test de rendimiento en comprensión lectora. La evaluación fue aplicada por los propios centros y profesores, que pudieron responder a las pruebas de manera convencional (papel y lápiz), o a través de la aplicación online.

En este capítulo, se lleva a cabo una descripción detallada de la muestra, además de los instrumentos utilizados para la recogida de datos y los procedimientos de análisis empleados.

6.1. Participantes

La población objeto de estudio corresponde a los estudiantes matriculados en 4º de Educación Primaria y 2º de Educación Secundaria Obligatoria de la Comunidad de Madrid.

En la tabla 6.1, se presenta el resumen de todos los datos disponibles en Primaria y en Secundaria. Tal y como se aprecia, hubo una preferencia entre la versión convencional, dado que un gran número de centros eligieron realizar la prueba por medio de papel y lápiz.

Tabla 6.1.
Resumen de los datos

	4º E.P		**2º E.S.O**	
	Papel	**Online**	**Papel**	**Online**
Comprensión lectora	9.258	1.079	46.482	2.207
Total	10.337		48.685	

Fuente: Elaboración propia

Para solucionar el evidente problema de sesgo debido a la no selección aleatoria (comentado en la parte teórica de la tesis), se procede a la utilización de la metodología de Puntaje de Propensión o Propensity Score, mediante el procedimiento de Matching

con el Vecino más cercano 1:10[9]. En la tabla 6.2, podemos observar la muestra balanceada con la que se van a realizar los estudios empíricos.

Tabla 6.2
Resumen de los datos después del Matching

	4º E.P		**2º E.S.O**	
	Papel	**Online**	**Papel**	**Online**
Comprensión lectora	5486	1.079	14511	2.207
Total	6565		16.714	

Fuente: Elaboración propia

Para conocer en detalle la estructura de nuestros datos, en la tabla 6.3 correspondiente a Primaria y la tabla 6.4 a Secundaria, se resumen las características de la muestra atendiendo a la titularidad de las escuelas y a las diferentes áreas territoriales.

Tabla 6.3
Resumen de los datos de Primaria en la prueba de Comprensión lectora por centros, estudiantes y áreas territoriales después del Matching

4º E.P		**Público**		**Privado Concertado**		**Privado**	
		Papel	Online	Papel	Online	Papel	Online
Centro	Frecuencias	767	117	978	131	345	134
	Porcentaje	14%	10,8%	17,8%	12,1%	6,3%	12,4%
Norte	Frecuencias	313	59	325	66	192	64
	Porcentaje	5,7%	5,5%	5,9%	6,1%	3,5%	5,9%
Sur	Frecuencias	1311	169	783	172	472	167
	Porcentaje	23,9%	15,7%	14,3%	15,9%	8,6%	15,5%
Este	Frecuencias	0	0	0	0	0	0
	Porcentaje	0%	0%	0%	0%	0%	0%
Total	Frecuencias	2391	345	2086	369	1009	365
	Porcentaje	43,6%	32,0%	38,0%	34,2%	18,4%	33,8%

Fuente: Elaboración propia

[9] Matching con el Vecino más cercano 1:10, es el procedimiento utilizando en este trabajo, por el que a cada sujeto que realiza la prueba online, se le asignan 10 sujetos que realizan la prueba en papel, y que tienen los valores más próximos en Propensity Score. Dicho procedimiento se detalla en el apartado 7.1 del capítulo de resultados de la tesis, donde se realiza la demostración del uso de las Propensity Score para el emparejamiento efectivo de muestras.

La muestra ha sido equilibrada en función del grupo tratado, es decir atendiendo al grupo que realiza la prueba online; lo que ha supuesto la no existencia de datos en los centros educativos de las áreas territoriales Este y Oeste, puesto que ningún centro de estas características realizó la prueba online. En Secundaria tampoco se dispone de sujetos en el distrito Sur de Madrid.

Tabla 6.4
Resumen de los datos de Secundaria en la prueba de Comprensión lectora por centros, estudiantes y áreas territoriales después del Matching

2ºE.S.O		**Público**		**Privado Concertado**		**Privado**	
		Papel	Online	Papel	Online	Papel	Online
Centro	Frecuencias	2121	256	5662	738	1524	381
	Porcentaje	14,6%	11,6%	39,0%	33,4%	10,5%	17,3%
Norte	Frecuencias	1195	212	607	142	696	133
	Porcentaje	8,2%	9,6%	4,2%	6,4%	4,8%	6,0%
Sur	Frecuencias	2706	345	0	0	0	0
	Porcentaje	18,6%	15,6%	0%	0%	0%	0%
Este	Frecuencias	0	0	0	0	0	0
	Porcentaje	0%	0%	0%	0%	0%	0%
Total	Frecuencias	6022	813	6269	880	2220	514
	Porcentaje	41,5%	36,8%	43,2%	39,9%	15,3%	23,3%

Fuente: Elaboración propia

6.2. Instrumentos

Las pruebas de Evaluación de Diagnóstico de 2010-2011 de la Comunidad de Madrid, son proporcionadas por la Consejería de Educación. Concretamente son tres pruebas de rendimiento: matemáticas, comprensión lectora y lengua, aplicadas a estudiantes de 2º Educación Secundaria Obligatoria y 4º de Educación Primaria de la Comunidad de Madrid[10]. En esta tesis doctoral se aborda la prueba de comprensión lectora en Primaria y Secundaria.

[10] Pueden verse las pruebas en la página de la Conserjería de Educación de la Comunidad de la Madrid, concretamente en el siguiente link: http://www.madrid.org/cs/Satellite?c=CM_InfPractica_FA&cid=1142502934515&idConsejeria=1109266187254&idListConsj=1109265444710&idOrganismo=1142359902140&language=es&pagename=ComunidadMadrid%2FEstructura&pid=1331802501654&pv=1142641697139&sm=1109170600517#links

Según recoge el B.O.C.M. Núm. 126 en el DECRETO 22/2007 y DECRETO 23/2007 del Consejo de Gobierno, por el que se establece para la Comunidad de Madrid el currículo de la Educación Primaria y de la Educación Secundaria Obligatoria, la educación literaria o comprensión lectora se entiende como *"el conjunto de habilidades y destrezas necesarias para leer de forma competente los textos literarios significativos de nuestro ámbito cultural"*.

La comprensión lectora es necesaria como materia instrumental en otros ámbitos no solo el escolar. Permite a los estudiantes obtener destrezas y habilidades para la comprensión de textos y las relaciones del texto literario con su contexto cultural. En el anexo 7, pueden observarse los objetivos de la enseñanza de la comprensión lectora para Primaria y Secundaria.

La prueba de comprensión lectora para Primaria y Secundaria (recopilada en los anexos 8 y 9 respectivamente) consta de 34 ítems de opción múltiple. En el caso de Secundaria se eliminó para los análisis el ítem 33 debido a sus problemas psicométricos (resumen en el anexo 10).

En la siguiente tabla 6.5, se resume el número y porcentaje de ítems atendiendo a los bloques de contenidos y a los procesos de la prueba de Comprensión lectora.

Tabla 6.5
Tabla de ponderaciones (número de ítems y porcentajes) de la prueba de Comprensión Lectora

		Procesos				
	Bloque de Contenidos	**Aproximación e identificación**	**Organización, síntesis e integración**	**Transferencia y aplicación**	**Reflexión y valoración**	**Total**
4º EP	Narrativos	3	2	2	2	9
	Descriptivos	2	2	0	4	9
	Expositivos	4	2	1	4	11
	Instructivos	3	1	0	1	6
	Total	12	7	4	11	34
	Porcentaje	35,6 %	20,5 %	12,3 %	31,5 %	100 %
2º ESO	Narrativos	2	1	1	1	7
	Descriptivos	2	2	0	3	7
	Expositivos	3	1	1	3	8
	Argumentativos	2	1	1	4	9
	Instructivos	2	1	0	1	4
	Total	11	7	4	12	34
	Porcentaje	33 %	20 %	13 %	43 %	100 %

Fuente: MESE (2010)

Los contenidos de aprendizaje, como se recoge en el DECRETO 22/2007 y 23/2007, "*requiere unos aprendizajes específicos que se inician en la educación Primaria con el recitado, la práctica de juegos retóricos, la escucha de textos propios de la literatura oral o las dramatizaciones. De este modo se consigue un primer acercamiento a las convenciones literarias básicas y a las relaciones entre las obras y el contexto histórico. Junto a todo ello, se favorecen experiencias placenteras con la lectura y la recreación de textos literarios. Esta orientación de la educación literaria continúa en la educación Secundaria obligatoria, de modo que se consolidan los hábitos de lectura, se amplían las experiencias en el campo de la lectura y recreación de textos, y se sistematizan las observaciones sobre las convenciones literarias y la relación entre las obras y sus contextos históricos*".

Basándonos en la clasificación de Werlich, las categorías de género textual utilizadas han sido: Textos Narrativos, Textos Descriptivos, Textos Expositivos, Textos Argumentativos y Textos Instructivos. En el anexo 11 se recogen cada una de estas categorías incluyendo los tipos de documentos tanto para Primaria como para Secundaria, así como los procesos vinculados con las destrezas de Comprensión lectora, que también se recogen en el anexo 12, tanto para Primaria como para Secundaria.

Conjuntamente, en el anexo 13 se presenta la matriz de especificaciones de Comprensión lectora con los descriptores para Primaria y en el anexo 14 para Secundaria, considerando los tipos de texto y los procesos descritos anteriormente.

6.3. Análisis estadísticos

Siguiendo las ideas recogidas por Muñiz y Hambleton (1999), para garantizar la equivalencia de ambas versiones (papel y online), es necesario que las medias, dispersiones y distribuciones de las puntuaciones sean aproximadamente las mismas. Unido al estudio del funcionamiento de los ítems para comprobar el comportamiento de cada uno de los ítems en ambas versiones de la prueba.

Para la evaluación de la equivalencia de ambas versiones, los análisis estadísticos utilizados han sido agrupados en cuatro grandes bloques:

Bloque I: Utilización de las Propensity Score para el emparejamiento efectivo de muestras y el posterior estudio de equivalencia y del Funcionamiento Diferencial de Versiones.

Para la demostración del primer objetivo se ha realizado un estudio comparativo de las diferentes técnicas Matching, por medio del paquete MatchIn del programa R (Ho, Imai y Stuart, 2011, 2015).

Bloque II: Validación según la Teoría Clásica de los Test.

Se realiza un análisis descriptivo y psicométrico con la finalidad de comprobar las medias, dispersiones, distribuciones de las puntuaciones; así como la fiabilidad en ambas versiones. De esta forma, se pondrá a prueba la primera hipótesis planteada y obtendremos una primera visión de ambas versiones. Lo que nos permitirá crear un punto de partida para continuar estudiando la equivalencia de las versiones. Estos análisis descriptivos han sido realizados con el software CORRECTOR 1.2-Complemento de MS-Excel, desarrollado por el Dr. JL Gaviria, del Dpto. de Métodos de Investigación y Diagnóstico en Educación de la Universidad Complutense de Madrid.

Bloque III: Validación según la Teoría de Respuesta al Ítem.

Es necesario conocer el modelo de la Teoría de Respuesta al Ítem que mejor ajusta en ambas versiones. Por ello, y para contrastar la segunda hipótesis planteada en este trabajo, se lleva a cabo un estudio comparativo del ajuste de nuestros datos atendiendo a modelos de 1 parámetro, 2 parámetros y 3 parámetros. Se ha utilizado el programa R, concretamente el paquete “difR”, versión 4.5 (Magis, et al. 2015).

Bloque IV: Estudio del supuesto de unidimensionalidad.

En este bloque de análisis se comprueba la tercera hipótesis sobre el estudio la invarianza factorial y por tanto el supuesto de unidimensionalidad de los datos. Aspecto relevante para los estudios de funcionamiento diferencial de los ítems y por tanto para la metodológica propuesta. Un ítem presenta indicios de DVF o DIF, cuando es multidimensional ya puede estar midiendo otros constructos además del factor principal. Schroeders y Wilhelm (2011) nos informan de la importancia y la utilidad de llevar a cabo análisis factoriales para realizar una correcta comparabilidad de los resultados entre ambas versiones. Se ha utilizado para ello el paquete "lavaan", versión 05-20 del programa R (Rosseel, Oberski, Byrnes, Vanbrabant, Savalei, Merkle, Hallquist, Rhemtulla, Katsikatsou y Barendse, 2015).

Bloque V: Análisis estadísticos para el estudio de la equivalencia de las versiones.

Conocida la estructura y las características de los datos, se procederá a través de análisis estadísticos a la verificación de la cuarta hipótesis relacionada con la comprobación de la equivalencia de ambas versiones. Para ello, se estudia la homogeneidad de varianzas, con la intención de evidenciar la existencia o no de asociación entre la puntuación en cada ítem y su modo de aplicación. Estudio realizado con el paquete SPSS, versión 22.

Concretamente con el análisis de homogeneidad de varianza detectamos si existen o no diferencias significativas en la distribución de las puntuaciones en función del modo de aplicación de la prueba. Observaremos si las varianzas y medias en ambas versiones son iguales o significativamente diferentes.

El estadístico Chi Cuadrado de Pearson nos permite poner a prueba la quinta hipótesis, estudiando si hay o no asociación entre la puntuación en cada ítem y el modo de aplicación. Para evitar el problema asociado a los tamaños muestrales del estadístico χ^2, se complementa esta prueba con medidas de asociación que no

se ven afectadas por el tamaño muestral, como son el Coeficiente de contingencia, V de Cramer y Phi (análisis realizados con el paquete SPSS, versión 22).

Bloque VI: Estudio del Funcionamiento Diferencial de Versiones.

Para finalizar y poner a prueba la sexta hipótesis planteada, llevamos a cabo un estudio aplicando la metodología propuesta en esta tesis: Funcionamiento Diferencial de Versiones.

Han sido utilizados varios procedimientos de detección de DVF con la intención de garantizar una mejor estimación y proporcionar conclusiones más consistentes y aproximadas.

Se recoge como análisis descriptivo algunos procedimientos basados en la Teoría Clásica de los Test (como el procedimiento T.I.D), otros procedimientos basados en la Teoría de Respuesta al Ítem (Rajú y Lord); y procedimientos basados en las tablas de contingencia (Estandarizado-Stand, Mantel – Haenszel).

Dichos métodos de detección presentan limitaciones considerables, principalmente basadas en su escasa potencia estadística para la detección de DIF (en unos casos el uniforme y en otros el no uniforme).

Por ello, y siguiendo a Swaminathan y Rogers (1990), se ha utilizado como alternativa la regresión logística (que analiza el DVF uniforme y no uniforme). Dicho procedimiento cuenta con una medida de magnitud (*pseudo-* R^2, con el procedimiento de McFadden) que nos permite conocer la magnitud del efecto para evitar los falsos DVF (Hidalgo, Gómez-Benito y Zumbo, 2014; Menard, 2000). Este es el motivo por el que profundizamos y aplicamos la Regresión Logística como método de detección del DVF. Se ha utilizado para su estimación el paquete "lordif" del programa R (Choi, Gibbons y Crane, 2011; 2015).

Para evitar problemas relativos al amplio tamaño muestral con el que se trabaja en este estudio, el nivel de significación establecido ha sido de 0,01. Pero además, en los análisis realizados para evaluar el Funcionamiento Diferencial de Versiones, se ajustaron los valores *p* con el método de Benjamin–Hochberg (1995) para lograr reducir la presencia de falsos positivos tan frecuentes en estudios con comparaciones múltiples.

CAPÍTULO 7: Resultados

"Dime y lo olvido, enséñame y lo recuerdo, involúcrame y lo aprendo"

Benjamin Franklin (1706- 1790)

7.1. Demostración del uso de las Propensity Score para el emparejamiento efectivo de muestras y el posterior estudio del Funcionamiento Diferencial de Versiones

En este apartado se expone el primer objetivo planteado en la tesis doctoral, correspondiente a la demostración del uso de las Propensity Score para el emparejamiento efectivo de muestras y para el posterior estudio del Funcionamiento Diferencial de Versiones.

Con la intención de realizar un análisis completo y garantizar la equivalencia entre los grupos, se ha realizado un estudio comparativo de las diferentes técnicas Matching (Vecino más cercano, Genético y Estratificación), para lo que se ha utilizado el paquete MatchIn del programa R (Ho, Imai y Stuart, 2015).

En la tabla 7.1 (Primaria) y 7.2 (Secundaria), se resumen los datos del modelo antes del Matching. En ambas tablas se observan las medias de la prueba online (grupo de control) y de la prueba en papel (grupo tratado).

Tabla 7.1
Resumen de los datos antes del Matching en Primaria

	Primaria			
	Media Online	Media Papel	Diferencia de Medias	Pr(>|t|)
	N= 1079	N= 9258		
distance	0,1403	0,1002	0,5353	0,000***
DAT	2,1168	2,0903	0,0294	0,000***
Tipo de Centro	2,0185	1,5891	0,5293	0,000***

Tabla 7.2
Resumen de los datos antes del Matching en Secundaria

	Secundaria			
	Media Online	Media Papel	Diferencia de Medias	Pr(>|t|)
	N= 2207	N= 46482		
distance	0,0641	0,0444	0,6976	0,000***
DAT	1,5333	2,413	-1,1737	0,000***
Tipo de Centro	1,8645	1,6342	0,3017	0,000***

En Primaria y Secundaria los resultados son contundentes, al mostrar la existencia significativa ($p<0,01$) de diferencias de medias entre la versión online y la versión en papel antes del matching.

En las tablas 7.3 y 7.4 pueden verse los resultados obtenidos en Primaria y Secundaria, respectivamente, tras el Matching. En el caso de la técnica "*Vecino más cercano"* se han llevado a cabo varias pruebas (K=2, K=4, K=5 y K=10), todas ellas con remplazo[11] (es decir, un sujeto que realiza la prueba online puede ser emparejado a varios sujetos que realizan la prueba en papel), y la probabilidad estimada por medio de Regresión Logística[12].

Tabla 7.3.
Estudio comparativo de las diferentes técnicas Matching en Primaria

		Matching				
		N Online	N Papel	Diferencia de medias	Tamaño del efecto	Pr(>|t|)
Vecino más cercano 2:1		1079	1807	0,0000	-1,9411	0,000***
Vecino más cercano 3:1		1079	2466	0,0000	-1,9411	0,000***
Vecino más cercano 4:1		1079	3104	0,0000	-1,8206	0,000***
Vecino más cercano 5:1		1079	3628	0,0000	-1,8852	0,000***
Vecino más cercano 10:1		**1079**	**5486**	**0,0000**	**-1,8913**	**0,000*** **
Genético		1079	993	0,0000	-1,7794	0,000***
Estratificado (5 estratos)	Estrato.1	176	1700	0,009	-	-
	Estrato .2	300	5198	0,022	-	-
	Estrato.3	66	412	0,000	-	-
	Estrato.4	306	1262	0,165	-	-
	Estrato.5	231	686	0,003	-	-
	ATE	-	-	-	-1,719	0,000***
	ATT	-	-	-	-1,933	0,000***

Fuente: Elaboración propia

Los resultados muestran un adecuado ajuste en todos los procedimientos, ya que las diferencias estandarizadas entre el grupo de control (online) y el grupo tratado

[11] Es posible realizar las estimaciones con o sin reemplazo, aunque en la mayoría de ocasiones se realiza con reemplazo, dado que reduce el sesgo, así como la precisión de los estimadores y la pérdida de numerosos datos (González, 2009). En cualquier caso, a la calidad de los resultados obtenidos no se vé condicionado ni afectado.

[12] También pueden utilizarse para estimar la probabilidad modelos probit o análisis discriminantes, pero son menos frecuentes.

(papel) son nulas, tomando en todos los casos el valor 0 (a excepción de alguno de los estratos en el procedimiento de estratificación).

Los resultados demuestran el acercamiento entre ambos grupos, logrando que desaparezcan las diferencias que existían entre ambos grupos antes del Matching.

En Secundaria (ver tabla 7.4) se observan resultados muy adecuados tras la realización del Matching.

Tabla 7.4
Estudio comparativo de las diferentes técnicas Matching en Secundaria

Matching						
		N Online	N Papel	Diferencia de medias	Tamaño del efecto	Pr(>|t|)
Vecino más cercano 2:1		2207	3994	0,0000	-0,9540	0,000***
Vecino más cercano 3:1		2207	5738	0,0000	-0,9540	0,000***
Vecino más cercano 4:1		2207	7278	0,0000	-0,8730	0,000***
Vecino más cercano 5:1		2207	8710	0,0000	-0,8807	0,000***
Vecino más cercano 10:1		**2207**	**14511**	**0,0000**	**-0.94628**	**0,000*** **
Genético		2207	2101	0,0000	-0,9166	0,000***
Estratificación (3 estratos)	Estrato.1	699	24601	0.3627	-	-
	Estrato .2	389	7427	0.0932	-	-
	Estrato .3	1119	7816	0.2387	-	-
	ATE	-	-	-	-0,423	0,000***
	ATT	-	-	-	-0,855	0,000***

Fuente: Elaboración propia

De nuevo, las diferencias estandarizadas entre los grupos (online y papel) son nulas en todos los procedimientos, con excepción del procedimiento de estratificación.

Una vez alcanzado el buen ajuste en ambas muestras debemos decantarnos por alguna de las técnicas presentadas para el análisis de nuestros datos. Por ello, la técnica seleccionada ha sido la del "*Vecino más cercano 1:10"*, dado que, además de tener un buen ajuste, la diferencia entre las medias de los dos grupos es 0 y, sobre todo, porque no perdemos un gran número de sujetos que realizan la prueba en papel y seguimos garantizando en todo momento la equivalencia.

Con esta técnica, a cada sujeto que realiza la prueba online se le empareja con 10 sujetos que realizan la prueba en papel con la misma puntuación Propensity Score.

A continuación se detallan los resultados obtenidos en Primaria (ver tabla 7.5):

Tabla 7.5
Resultado Propensity Score técnica Matching: Vecino más cercano 10:1 con remplazo "double robustness" en Primaria

Resumen del equilibrio con todos los datos:

	Media Online	Media Papel	SD Informatizado	Diferencia de Medias	eCDF Med	eCDF Media	eCDF Max
distance	0,1403	0,1002	0,0562	0,5353	0,1400	0,1461	0,3039
DAT	2,1168	2,0903	0,9396	0,0294	0,0198	0,0220	0,0463
Tipo de Centro	2,0185	1,5891	0,6891	0,5293	0,2026	0,1431	0,2268

Resumen del equilibrio para los datos emparejados - Matching:

	Media Online	Media Papel	SD Informatizado	Diferencia de Medias	eCDF Med	eCDF Media	eCDF Max
distance	0,1403	0,1403	0.0749	0,0000	0.0931	0,0935	0,1730
DAT	2,1168	2,1168	0.9007	0,0000	0.0031	0,0100	0,0269
Tipo de Centro	2,0185	2,0185	0.8110	0,0000	0.1161	0,0901	0,1544

Porcentaje de mejora del emparejamiento - Matching:

	Std. Mean Diff.	eCDF Med	eCDF Mean	eCDF Max
distance	100	33.5293	36.0060	43.0935
DAT	100	84.4909	54.5790	41.7815
TIPOCENTRO	100	42.7021	37.0203	31.9444

Tamaño de la muestra:

	Papel	Online
All	9258	1079
Matched	5486	1079
Unmatched	3772	0
Discarded	0	0

Fuente: Elaboración propia

La tabla 7.5 nos presenta de nuevo se indica un resumen de todos los datos antes de realizar el Matching y otros sobre el equilibrio con los datos emparejados tras el Matching. Donde se aprecia como no existe diferencia de medias, tal y como se ha mencionado anteriormente.

En el siguiente apartado de la tabla 7.5, se lleva a cabo un análisis porcentual de la mejora del emparejamiento tras el Matching, donde se demuestra que en todas las variables ha sido posible encontrar al vecino más cercano en el 100% de los casos.

Por ese motivo, la muestra se compone de todos los sujetos del grupo de control (online = 1.079) y una disminución de sujetos (concretamente de 3.772) en el grupo tratado (papel = 5.486).

En lo que respecta a Secundaria, los resultados pueden observarse en la tabla 7.6.

Tabla 7.6.
Resultado Propensity Score técnica Matching: Vecino más cercano 10:1 con remplazo "double robustness" en Secundaria

Resumen del equilibrio con todos los datos:

	Media Online	Media Papel	SD Informatizado	Diferencia de Medias	eCDF Med	eCDF Media	eCDF Max
Apldistance	0,0641	0,0444	0,0289	0,6976	0,2252	0,1937	0,3069
DAT	1,5333	2,4130	0,4761	-1,1737	0,1725	0,1759	0,3216
Tipo de Centro	1,8645	1,6342	0,6668	0,3017	0,1038	0,0768	0,1266

Resumen del equilibrio para los datos emparejados - Matching:

	Media Online	Media Papel	SD Informatizado	Diferencia de Medias	eCDF Med	eCDF Media	eCDF Max
Ap distance	0,0641	0,0641	0,0282	0	0,0165	0,0223	0,0676
DAT	1,5333	1,5333	0,7494	0	0,0184	0,0162	0,0302
Tipo de Centro	1,8645	1,8645	0,7635	0	0,0466	0,0422	0,0799

Porcentaje de mejora del emparejamiento - Matching:

	Std. Mean Diff.	eCDF Med	eCDF Mean	eCDF Max
distance	100	99,6948	88,4782	77,9729
DAT	100	89,3604	90,8084	90,6226
TIPOCENTRO	100	55,0701	45,0735	36,8796

Tamaño de la muestra:

	Papel	Online
All	46482	2207
Matched	14511	2207
Unmatched	31971	0
Discarded	0	0

Fuente: Elaboración propia

De nuevo se aprecia la existencia de diferencia de medias antes de llevar a cabo el emparejamiento; pero una vez que la muestra ha sido balanceada, la diferencia de medias es nula.

El acercamiento del grupo de control al grupo de tratamiento es muy relevante; las dos covariables introducidas en el modelo se pudieron encontrar con el vecino más cercano en el 100% de los casos.

También podemos observar cómo ha variado el tamaño de la muestra: una vez que los datos han sido emparejados tienden a igualarse el grupo de control (online = 2.207) con el grupo tratado (papel = 14.511), para lo cual ha sido necesario la eliminación de un volumen importante de sujetos tratados, concretamente de 31.971.

Esta equivalencia puede verse de forma gráfica en las figuras 9 (para Primaria) y 10 (para Secundaria).

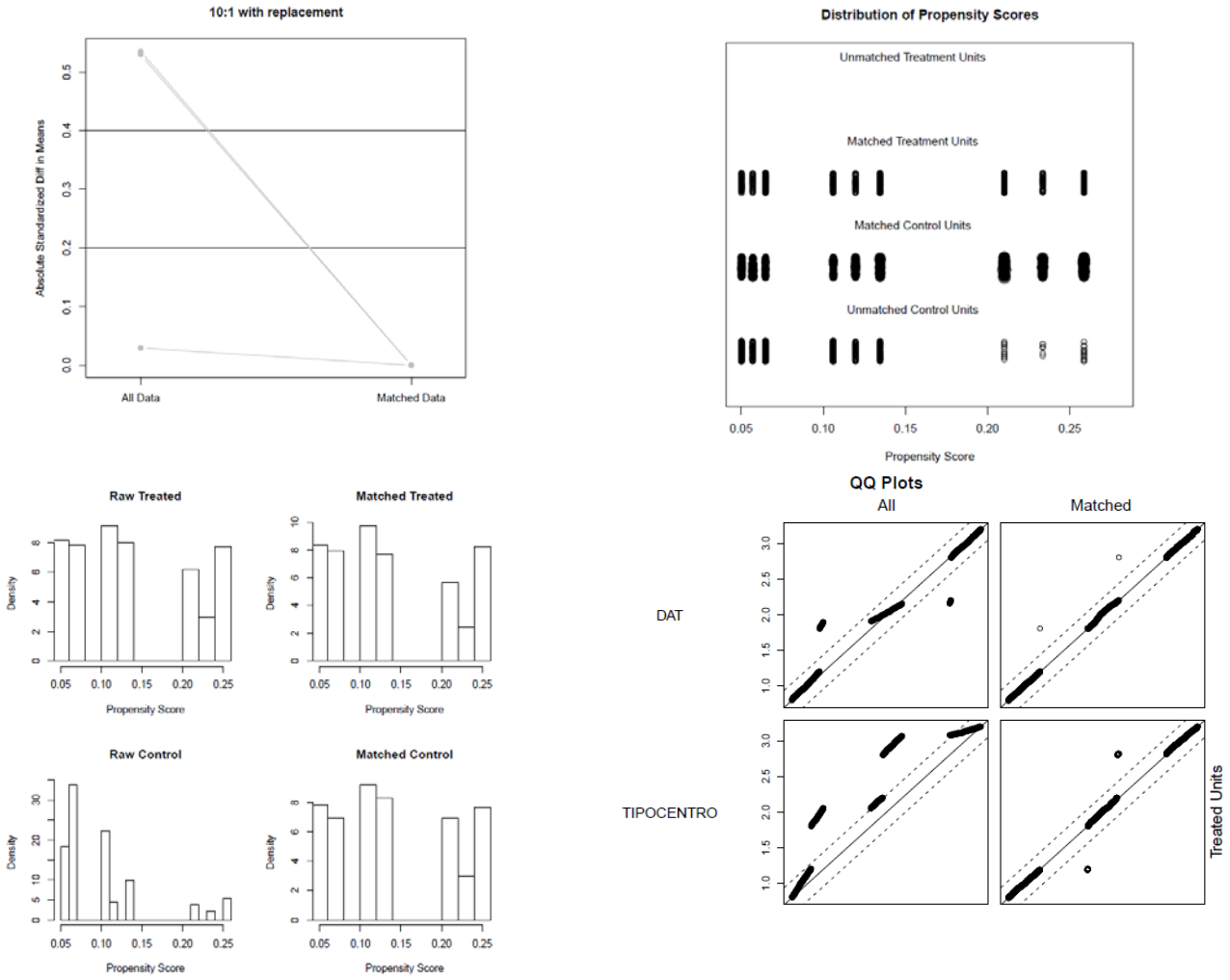


Figura 9. Resultado Propensity Score técnica Matching: Vecino más cercano 10:1 con remplazo "double robustness" en Primaria.

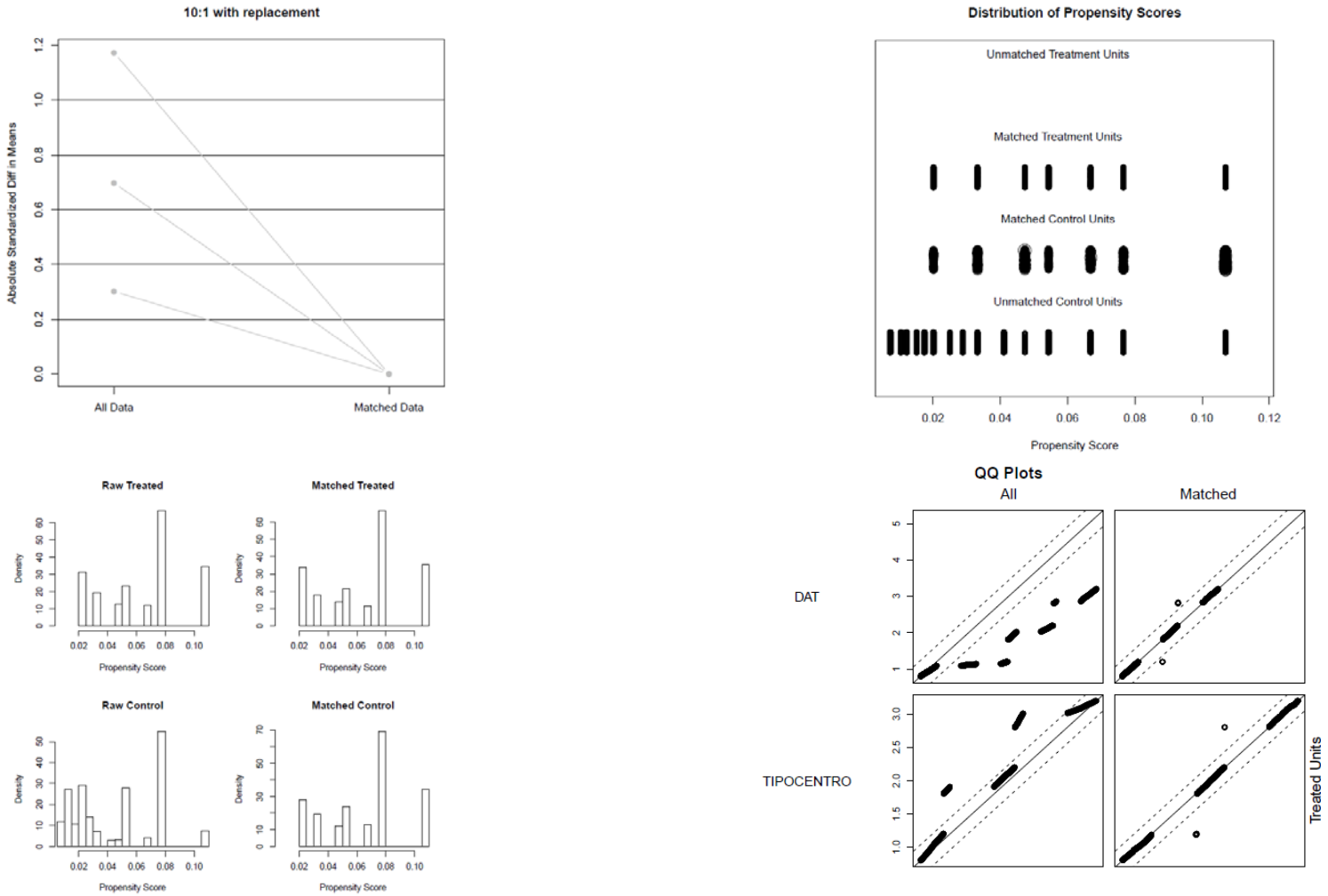


Figura 10. Resultado Propensity Score técnica Matching: Vecino más cercano 10:1 con remplazo "double robustness" en Secundaria

Para finalizar este apartado es esencial el estudio del efecto del tratamiento[13], para lo cual se lleva a cabo una regresión múltiple, cuya variable dependiente es la puntuación en la prueba y como variables independientes las covariables (Distrito y Tipo de centro).

En las tablas 7.7 (para Primaria) y 7.8 (para Secundaria) podemos observar los resultados obtenidos. En ambos casos, el modo de aplicación, es decir, el efecto del tratamiento, es significativo.

Tabla 7.7
Efecto del tratamiento en la variable Puntuación Total (Treatment Effect Estimation) en Primaria

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.24870	0.24535	86.607	< 2e-16 ***
APLICACION	-1.89132	0.18869	-10.023	< 2e-16 ***
DAT	0.13491	0.07598	1.776	0.0758 .
TIPOCENTRO	0.39313	0.09179	4.283	1.87e-05 ***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.616 on 6561 degrees of freedom
Multiple R-squared: 0.01667, Adjusted R-squared: 0.01622
F-statistic: 37.07 on 3 and 6561 DF, p-value: < 2.2e-16

Evaluación en 1000 simulaciones

Model: ls
Number of simulations: 1000

Expected Values: E(Y|X)

Mean	sd	50%	2.5%	97.5%
21.88	0.076	21.87	21.73	22.03

First Differences: E(Y|X1) - E(Y|X)

Mean	sd	50%	2.5%	97.5%
0	0	0	0	0

[13] Se utiliza la función "Zelig" que realiza una evaluación en 1000 simulaciones.

Tabla 7.8
Efecto del tratamiento en la variable Puntuación Total (Treatment Effect Estimation). Secundaria

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20,58184	0.1723	114,836	< 2e-16 ***
APLICACION	-0,94628	0.1194	-7,923	2,47e-15 ***
DAT	0.36023	0.05858	6,150	7,93e-10 ***
TIPOCENTRO	1,70068	0.06420	0,06420	< 2e-16 ***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.216 on 16714 degrees of freedom
Multiple R-squared: 0.04533, Adjusted R-squared: 0.04516
F-statistic: 264.5 on 3 and 16714 DF, p-value: < 2.2e-16

Evaluación en 1000 simulaciones

Model: ls
Number of simulations: 1000

Expected Values: E(Y|X)					First Differences: E(Y|X1) - E(Y|X)				
Mean	sd	50%	2.5%	97.5%	Mean	sd	50%	2.5%	97.5%
23.55	0.042	23.55	23.46	22.63	0	0	0	0	0

Al llevar a cabo las 1.000 simulaciones, la diferencia de los valores esperados de la puntuación del grupo de control con el grupo de tratamiento es de 0, lo cual demuestra que no existe ningún efecto del tratamiento en la puntuación total (tal y como se verá en apartado 7.5. Análisis estadísticos para el estudio de la equivalencia).

Verificada la equivalencia de ambos grupos de sujetos, es posible poner en práctica la metodología propuesta en este trabajo, “Funcionamiento Diferencial de Versiones”, y poder analizar la equivalencia entre ambas versiones.

7.2. Validación según la Teoría Clásica de los Test de la prueba de Comprensión Lectora en Primaria y Secundaria

A continuación, presentamos un análisis descriptivo de la prueba de Comprensión lectora. En la tabla 7.9, resumimos algunos descriptivos en función del modo de aplicación y atendiendo a la muestra balanceada con los valores del Propensity Score.

Tabla 7.9.
Resumen descriptivo de la prueba de Comprensión Lectora

	Primaria		Secundaria	
	Papel	Online	Papel	Online
Nº de ítems	34	34	33	33
Nº de sujetos	5486	1079	14511	2207
Mínimo	4,00	5,00	2,00	0,00
Máximo	34,00	34,00	33,00	33,00
Nº medio de aciertos	22,31	20,44	23,84	24,33
Error típico de la media	0,07	0,17	0,04	0,10
Varianza	28,40	32,26	26,60	21,05
Desv. típica	5,33	5,68	5,16	4,59
Facilidad media	0,66	0,60	0,72	0,74
Rbp media	0,36	0,38	0,38	0,35

Fuente: Elaboración propia

Las características métricas en la prueba de Comprensión lectora en Primaria, arrojan resultados semejantes, pero siempre superiores en la prueba en papel (mayor número de aciertos y mayor facilidad), mientras que la correlación biserial puntual, varianza y desviación típica son ligeramente superiores en la prueba online.

En Secundaria también se alcanzan valores semejantes, pero en este caso superiores en la prueba online. Puede observarse como es mayor el número de aciertos y la facilidad en la prueba online, mientras que la correlación biserial puntual, varianza y desviación típica, aunque semejantes, son superiores en la prueba en papel. Estos

resultados aportan evidencias a favor de la equivalencia en cuanto a las propiedades psicométricas se refiere, arrojando en ambas versiones comportamientos semejantes. En los anexos 15 y 16, se detallan los estadísticos descriptivos para cada uno de los ítems en ambas versiones en Primaria y en los anexos 17 y 18 para Secundaria.

A continuación, nos centraremos exclusivamente en la puntuación total de la prueba, pudiendo observar los resultados en la tabla 7.10.

Tabla 7.10.
Estadístico descriptivo de la puntuación en la prueba de Comprensión Lectora

Puntuación total		N	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
						Estad.	Error típ	Estad.	Error típ
Primaria	Papel	5486	22,22	5,61	31,49	-,400	,033	-,330	,066
	Online	1079	20,44	5,68	32,29	-,298	,074	-,440	,149
Secundaria	Papel	14511	24,09	5,30	28,14	-1,20	0,02	1,701	0,04
	Online	2207	23,36	5,51	30,42	-0,86	0,05	0,328	0,10

Fuente: Elaboración propia

Los valores en la puntuación total en ambas versiones son semejantes. Se podría destacar una ligera diferencia a favor de la versión en papel, tanto en Primaria como Secundaria, no siendo superior a 2 puntos de diferencia.

En la figura 11 (Primaria) y figura 12 (Secundaria), se observa la misma distribución, dado que las frecuencias de las puntuaciones en la versión en papel y en la versión online son muy similares. En Primaria, la puntuación total toma valores centrales, mientras que en Secundaria se obtienen puntuaciones mayores.

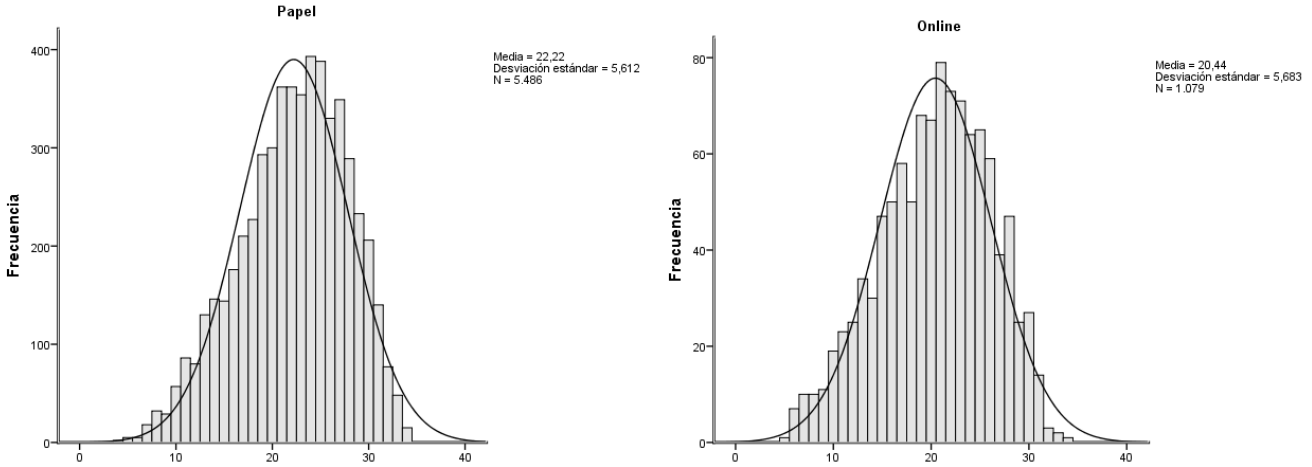


Figura 11. Frecuencia de las puntuaciones en la versión en papel y online de Primaria

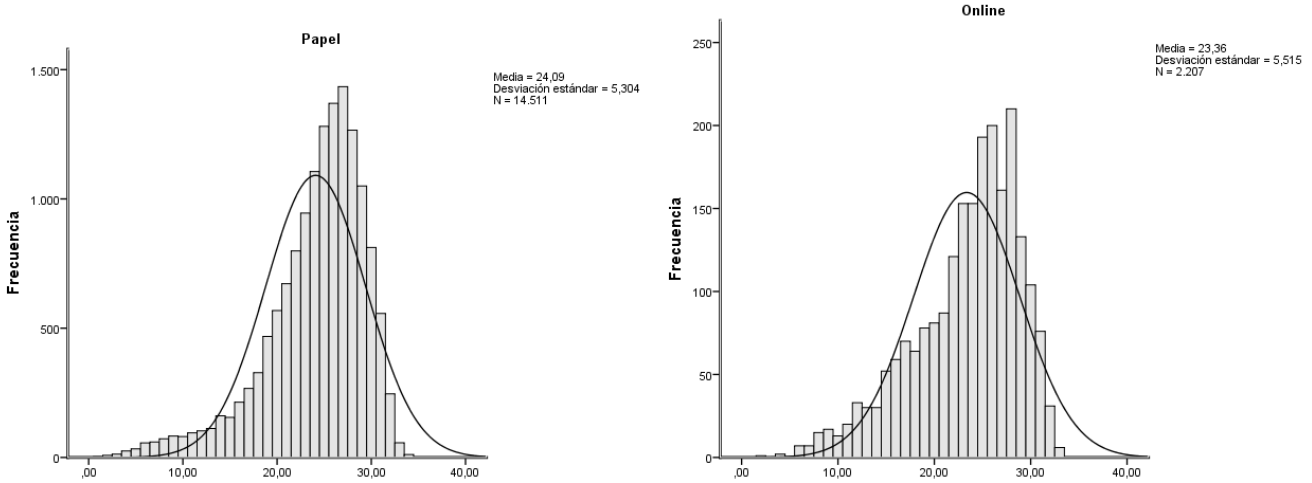


Figura 12. Frecuencia de las puntuaciones en la versión en papel y online de Secundaria

Las características de cada uno de los ítems, atendiendo a la Teoría Clásica de los Test, son recogidas en los anexos 19 y 20 (Primaria) y 21 y 22 (Secundaria), donde de nuevo se puede observar la semejanza en el comportamiento de los ítems atendiendo al modo de aplicación de la prueba.

Confiabilidad por consistencia interna

Para comprobar la equivalencia entre las dos versiones, es requisito necesario estudiar la fiabilidad de ellas. Es imprescindible que ambas versiones tengan índices equivalentes. A continuación, en la tabla 7.11 se recoge el coeficiente de fiabilidad Alfa de Cronbach para ambas versiones:

Tabla 7.11.
Análisis de fiabilidad en la prueba de Comprensión Lectora

		N casos	Alfa de Cronbach	N de ítems
Primaria	Papel	5486	0,812 (0,000)	34
	Online	1079	0,811 (0,000)	34
Secundaria	Papel	14511	0,837 (0,000)	33
	Online	2207	0,830 (0,000)	33

Fuente: Elaboración propia

Podemos observar que en la prueba de Comprensión lectora, el coeficiente de fiabilidad en Primaria es ligeramente superior en la versión en papel que en la versión online; concretamente con una fiabilidad significativa de 0,812 en papel y 0,811 en online. Lo mismo sucede en Secundaria, la fiabilidad es algo superior en papel (0,837) que en la versión online (0,830), pero de nuevo significativa. Los resultados son contundentes al mostrar que ambos coeficientes de fiabilidad, en Primaria y Secundaria, son equivalentes y adecuados.

Comprobada la primera hipótesis, los resultados confirman la no existencia de diferencias entre ambas versiones (papel y lápiz – online) de la prueba de rendimiento en comprensión lectora; en lo que a las características psicométricas de la Teoría Clásica de los Test se refiere.

7.3. Validación según la Teoría de Respuesta al Ítem

Los parámetros de los modelos de TRI, son estimados con el paquete difR, concretamente con la función itemParEst (Item parameter estimation for DIF detection), muy apropiada para nuestro estudio en el que posteriormente llevaremos a cabo estudios de detección del DVF. Se estiman estos parámetros atendiendo a un modelo de regresión logística[14.]

En los anexos 23–28 (Primaria) y 29-34 (Secundaria), podemos observar el resultado de los tres modelos llevados a cabo en ambas versiones, junto a los parámetros [a, b, c], los errores asociados a cada parámetro [se (a), se (b), se (c)] y las covarianzas entre los parámetros [cov(a,b), cov(a,c), cov(b,c)]; además de estadísticos de ajuste de los ítems al modelo (Chi cuadrado).

Ambas versiones tienen un ajuste muy adecuado en el modelo de uno, dos y tres parámetros; pero el modelo que mejor ajusta es el de dos parámetros, tanto en Primaria como en Secundaria. En dicho modelo tan solo dos ítems no ajustan en la versión en papel (ver tabla 7.12) y cuatro ítems no ajustan en la versión online (ver tabla 7.13).

[14] El modelo de 1 parámetro es estimado utilizando un modelo mixto generalizado lineal (generalized linear mixed model), mientras que los modelos de 2 y 3 parámetros se estiman a través del método de máxima verosimilitud.

Tabla 7.12.
Modelo de 2 Parámetro en la versión en papel de Primaria

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,805	-1,071	0,090	0,131	0,008	5,660	7,000	0,580
Ítem2	0,828	-2,463	0,113	0,291	0,030	5,074	7,000	0,651
Ítem3	1,728	-1,791	0,183	0,122	0,018	5,003	7,000	0,660
Ítem4	0,706	0,339	0,083	0,102	-0,003	22,888	7,000	**0,002**
Ítem5	0,392	0,804	0,073	0,216	-0,010	9,771	7,000	0,202
Ítem6	0,795	-1,087	0,090	0,134	0,009	12,978	7,000	0,073
Ítem7	0,872	-2,317	0,114	0,259	0,027	17,565	7,000	0,014
Ítem8	0,763	-0,736	0,086	0,114	0,006	15,499	7,000	0,030
Ítem9	1,008	-2,661	0,140	0,298	0,039	7,987	7,000	0,334
Ítem10	0,553	0,425	0,078	0,130	-0,004	15,079	7,000	0,035
Ítem11	0,365	1,465	0,074	0,334	-0,021	12,009	7,000	0,100
Ítem12	0,753	0,030	0,084	0,091	0,000	14,797	7,000	0,039
Ítem13	0,774	0,572	0,087	0,104	-0,004	23,951	7,000	**0,001**
Ítem14	0,762	0,019	0,084	0,091	0,000	18,261	7,000	0,011
Ítem15	1,236	-1,559	0,125	0,128	0,013	6,365	7,000	0,498
Ítem16	1,107	-1,419	0,112	0,126	0,011	6,032	7,000	0,536
Ítem17	0,925	-0,282	0,092	0,081	0,002	13,042	7,000	0,071
Ítem18	0,887	-0,604	0,091	0,094	0,004	9,662	7,000	0,209
Ítem19	0,462	0,742	0,075	0,179	-0,008	13,010	7,000	0,072
Ítem20	0,842	0,730	0,090	0,106	-0,005	3,961	7,000	1,497
Ítem21	0,375	-0,353	0,072	0,180	0,005	8,116	7,000	0,322
Ítem22	1,421	-1,343	0,134	0,101	0,010	7,088	7,000	0,420
Ítem23	1,324	-1,773	0,139	0,141	0,016	10,779	7,000	0,149
Ítem24	0,657	-1,391	0,085	0,187	0,013	7,528	7,000	0,376
Ítem25	0,635	-0,100	0,080	0,106	0,001	8,691	7,000	0,276
Ítem26	1,388	-1,101	0,126	0,089	0,007	11,060	7,000	0,136
Ítem27	1,730	-1,579	0,172	0,105	0,014	12,885	7,000	0,075
Ítem28	2,102	-1,020	0,187	0,067	0,007	5,348	7,000	0,618
Ítem29	1,077	-1,244	0,107	0,115	0,009	12,266	7,000	0,092
Ítem30	0,857	-1,331	0,095	0,145	0,011	8,848	7,000	0,264
Ítem31	1,254	-1,972	0,140	0,166	0,020	5,039	7,000	0,655
Ítem32	0,495	-0,985	0,077	0,191	0,010	5,954	7,000	0,545
Ítem33	0,980	-0,008	0,094	0,075	0,000	17,348	7,000	0,015
Ítem34	0,933	-1,722	0,106	0,172	0,015	11,602	7,000	0,114

Fuente: Elaboración propia
Nota: en negrita se indican los ítems que no presentan un ajuste adecuado (p<0,01)

Tabla 7.13.
Modelo de 2 Parámetro en la versión online de Primaria

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,755	-0,502	0,084	0,101	0,004	10,892	7,000	0,143
Ítem2	0,928	-1,979	0,114	0,207	0,021	6,004	7,000	0,539
Ítem3	1,644	-1,533	0,166	0,107	0,014	10,374	7,000	0,168
Ítem4	0,789	1,159	0,088	0,140	-0,009	15,220	7,000	0,033
Ítem5	0,294	1,533	0,071	0,416	-0,025	4,502	7,000	0,721
Ítem6	0,942	-0,702	0,094	0,093	0,005	7,406	7,000	0,388
Ítem7	1,108	-1,527	0,117	0,136	0,013	11,672	7,000	0,112
Ítem8	0,843	-0,339	0,087	0,087	0,002	17,240	7,000	0,016
Ítem9	1,238	-1,904	0,142	0,166	0,020	4,626	7,000	0,705
Ítem10	0,637	1,102	0,081	0,162	-0,010	3,248	7,000	**0,000**
Ítem11	0,104	9,762	0,076	7,121	-0,537	9,051	7,000	0,249
Ítem12	0,603	0,258	0,077	0,114	-0,002	3,798	7,000	0,803
Ítem13	0,498	1,634	0,078	0,269	-0,018	3,832	7,000	2,632
Ítem14	0,839	0,467	0,086	0,093	-0,003	9,170	7,000	0,241
Ítem15	1,220	-1,545	0,127	0,129	0,013	5,367	7,000	0,615
Ítem16	1,016	-1,015	0,101	0,104	0,007	13,216	7,000	0,067
Ítem17	1,137	0,157	0,100	0,069	-0,001	10,835	7,000	0,146
Ítem18	0,774	-0,278	0,084	0,092	0,002	10,968	7,000	0,140
Ítem19	0,337	1,978	0,073	0,453	-0,030	2,961	7,000	0,889
Ítem20	0,674	1,656	0,086	0,211	-0,015	28,610	7,000	**0,000**
Ítem21	0,463	-0,346	0,074	0,147	0,004	7,687	7,000	0,361
Ítem22	1,476	-0,954	0,131	0,077	0,006	19,501	7,000	**0,007**
Ítem23	2,350	-1,383	0,242	0,079	0,013	3,542	7,000	0,831
Ítem24	0,575	-1,949	0,087	0,287	0,022	18,444	7,000	0,010
Ítem25	0,564	0,352	0,076	0,124	-0,003	10,905	7,000	0,143
Ítem26	1,198	-0,913	0,111	0,087	0,006	6,028	7,000	0,536
Ítem27	1,936	-1,308	0,185	0,082	0,010	7,237	7,000	0,405
Ítem28	1,954	-0,725	0,164	0,058	0,004	10,987	7,000	0,139
Ítem29	1,179	-0,755	0,107	0,080	0,005	9,176	7,000	0,240
Ítem30	0,948	-0,835	0,095	0,099	0,006	13,743	7,000	0,056
Ítem31	1,454	-1,769	0,158	0,137	0,018	4,555	7,000	0,714
Ítem32	0,516	-0,828	0,077	0,167	0,008	19,055	7,000	**0,008**
Ítem33	1,084	0,031	0,098	0,070	0,000	11,331	7,000	0,125
Ítem34	1,183	-1,658	0,127	0,143	0,015	4,612	7,000	0,707

Fuente: Elaboración propia
Nota: en negrita se indican los ítems que no presentan un ajuste adecuado (p<0,01)

En Secundaria sucede lo mismo, el mejor ajuste se da en el modelo de 2 parámetros. En la versión en papel, (ver tabla 7.14) no ajustan cinco ítems y en la versión online, (ver tabla 7.15) no ajustan dos ítems.

Tabla 7.14.
Modelo de 2 Parámetro en la versión en papel de Secundaria

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,325	-3,195	0,082	0,798	0,063	4,312	7,000	0,743
Ítem2	1,227	-2,208	0,141	0,197	0,024	6,825	7,000	0,447
Ítem3	0,939	-2,342	0,118	0,252	0,027	9,151	7,000	0,242
Ítem4	0,831	-1,872	0,103	0,213	0,019	5,263	7,000	0,628
Ítem5	1,512	-1,827	0,159	0,141	0,018	3,800	7,000	0,803
Ítem6	0,319	-3,089	0,081	0,784	0,061	15,224	7,000	0,033
Ítem7	0,616	-0,910	0,085	0,159	0,009	22,914	7,000	**0,002**
Ítem8	0,467	-0,087	0,079	0,143	0,001	8,256	7,000	0,311
Ítem9	0,350	0,854	0,077	0,258	-0,014	10,423	7,000	0,166
Ítem10	0,587	-0,566	0,083	0,138	0,006	20,772	7,000	**0,004**
Ítem11	0,774	-2,827	0,113	0,365	0,039	6,847	7,000	0,445
Ítem12	1,155	-1,508	0,121	0,138	0,013	5,056	7,000	0,653
Ítem13	1,853	-1,603	0,187	0,110	0,015	6,043	7,000	0,535
Ítem14	1,229	-0,702	0,120	0,084	0,006	10,642	7,000	0,155
Ítem15	0,805	-0,873	0,094	0,125	0,008	11,402	7,000	0,122
Ítem16	0,658	-1,218	0,088	0,179	0,012	6,799	7,000	0,450
Ítem17	1,078	0,153	0,110	0,073	-0,001	26,111	7,000	**0,000**
Ítem18	1,399	-1,481	0,140	0,120	0,013	11,220	7,000	0,129
Ítem19	0,290	1,844	0,078	0,528	-0,037	10,481	7,000	0,163
Ítem20	1,284	-0,691	0,123	0,081	0,005	12,367	7,000	0,089
Ítem21	1,617	-1,642	0,164	0,121	0,015	3,539	7,000	0,831
Ítem22	0,910	-1,133	0,101	0,131	0,010	22,176	7,000	**0,002**
Ítem23	1,667	-1,815	0,174	0,132	0,018	6,344	7,000	0,500
Ítem24	1,265	-2,064	0,141	0,179	0,022	41,569	7,000	**0,000**
Ítem25	0,709	-1,179	0,090	0,164	0,011	10,645	7,000	0,155
Ítem26	1,403	-1,622	0,143	0,130	0,015	4,806	7,000	0,684
Ítem27	1,684	-1,333	0,163	0,099	0,012	13,690	7,000	0,057
Ítem28	1,395	-1,271	0,137	0,107	0,011	5,841	7,000	0,558
Ítem29	1,084	-1,206	0,113	0,121	0,010	11,195	7,000	0,130
Ítem30	1,498	-2,182	0,168	0,173	0,024	11,993	7,000	0,101
Ítem31	1,781	-1,889	0,190	0,133	0,020	4,273	7,000	0,748
Ítem32	1,052	-1,132	0,110	0,118	0,010	5,292	7,000	0,624
Ítem34	0,325	-3,195	0,082	0,798	0,063	4,312	7,000	0,743

Fuente: Elaboración propia
Nota: en negrita se indican los ítems que no presentan un ajuste adecuado (p<0,01)

Tabla 7.15.
Modelo de 2 Parámetro en la versión en online de Secundaria

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,276	-3,553	0,080	1,033	0,080	7,462	7,000	0,382
Ítem2	1,015	-2,420	0,132	0,256	0,030	7,553	7,000	0,374
Ítem3	0,681	-2,156	0,097	0,287	0,025	21,681	7,000	**0,003**
Ítem4	1,012	-1,324	0,106	0,132	0,011	6,466	7,000	0,487
Ítem5	1,256	-2,152	0,148	0,191	0,024	4,392	7,000	0,734
Ítem6	0,285	-3,531	0,081	0,999	0,078	6,988	7,000	0,430
Ítem7	0,645	-0,626	0,083	0,129	0,006	12,225	7,000	0,093
Ítem8	0,691	-0,104	0,085	0,102	0,001	11,098	7,000	0,134
Ítem9	0,085	7,576	0,075	6,679	-0,494	13,409	7,000	0,063
Ítem10	0,611	-0,309	0,082	0,119	0,003	14,162	7,000	0,048
Ítem11	0,602	-2,808	0,100	0,433	0,041	13,705	7,000	0,057
Ítem12	1,239	-0,733	0,114	0,082	0,005	11,444	7,000	0,120
Ítem13	1,364	-1,524	0,136	0,120	0,013	4,380	7,000	0,735
Ítem14	1,204	-0,292	0,110	0,071	0,002	14,402	7,000	0,044
Ítem15	0,908	-0,055	0,094	0,082	0,001	11,317	7,000	0,125
Ítem16	0,717	-1,375	0,090	0,179	0,013	15,419	7,000	0,031
Ítem17	0,804	0,211	0,090	0,092	-0,002	11,694	7,000	0,111
Ítem18	1,283	-1,063	0,120	0,096	0,008	10,210	7,000	0,177
Ítem19	0,286	2,703	0,079	0,756	-0,057	6,548	7,000	0,477
Ítem20	1,203	-0,888	0,112	0,090	0,006	18,165	7,000	0,112
Ítem21	1,939	-0,984	0,171	0,072	0,007	8,465	7,000	0,293
Ítem22	0,883	-0,317	0,093	0,088	0,002	12,681	7,000	0,080
Ítem23	1,718	-1,519	0,168	0,105	0,013	4,859	7,000	0,677
Ítem24	1,081	-2,508	0,143	0,262	0,034	18,553	7,000	**0,010**
Ítem25	0,650	-0,727	0,084	0,135	0,007	16,627	7,000	0,020
Ítem26	1,590	-0,988	0,141	0,080	0,007	10,693	7,000	0,153
Ítem27	1,591	-1,115	0,145	0,086	0,008	13,964	7,000	0,052
Ítem28	1,290	-1,182	0,123	0,102	0,009	7,907	7,000	0,341
Ítem29	1,012	-1,124	0,104	0,117	0,009	7,599	7,000	0,369
Ítem30	2,116	-1,827	0,239	0,116	0,021	7,838	7,000	0,347
Ítem31	1,752	-1,656	0,179	0,114	0,015	8,899	7,000	0,260
Ítem32	0,868	-1,181	0,096	0,136	0,010	7,818	7,000	0,349
Ítem34	0,656	-0,814	0,084	0,141	0,008	5,374	7,000	0,614

Fuente: Elaboración propia
Nota: en negrita se indican los ítems que no presentan un ajuste adecuado (p<0,01)

Los resultados obtenidos aportan evidencias a favor del correcto ajuste en el modelo de dos parámetros en ambas versiones. En virtud de su buen ajuste, los análisis presentados posteriormente atenderán a dicho modelo. Asimismo, los análisis del Funcionamiento Diferencial de Versiones realizan la estimación considerando este modelo como el más oportuno.

7.4. Estudio del supuesto de la unidimensionalidad

Cuando realizamos estudios DIF, o en nuestro caso DVF, es imprescindible estudiar el supuesto de unidimensionalidad.

Según Ackerman (1992) y Herrera (2005) un ítem no presenta DIF cuando es unidimensional o cuando la distribución de la dimensión irrelevante en los grupos comparados y los parámetros TRI son los mismos. Por lo tanto, un ítem presenta DIF cuando está midiendo además de la dimensión principal (θ), un constructo secundario o una dimensión irrelevante (η) y cuando las varianzas en las distribuciones de la dimensión irrelevante, fijado el valor θ, son diferentes en ambos grupos comparados (Bandeira, 2002).

Por ello, es de vital relevancia llevar a cabo un estudio de la unidimensionalidad que nos permita anticipar si estamos ante ítems sin DVF o con DVF dada la multidimensionalidad, ya que puede ser un factor que nos de indicios de estar midiendo otros constructos latentes, además del factor principal.

Dado que los ítems han sido recodificados como 1 acierto y 0 error, estamos ante variables de carácter dicotómico, por lo que para el estudio de la unidimensionalidad de las versiones, así como la estimación de los parámetros e índices de ajuste, se ha utilizado la matriz de correlaciones tetracóricas y el método de estimación de Mínimos Cuadrados Ponderados Diagonalizados (DWLS, Diagonal Weighted Least Squares)[15] (Mîndrilã, 2010).

Para este estudio, se ha realizado un análisis factorial confirmatorio con un modelo donde todos los ítems pesan en un único factor, tanto para la muestra en papel, como la online. Para la estimación de los modelos, se fijó en 1 las varianzas de las variables latentes. De manera resumida se presentan en la tabla 7.16 los resultados obtenidos.

[15] Se recomienda su versión robusta porque ofrece mayores tasas de convergencia. Además es muy apropiado con datos dicotómicos, porque con muestras grandes como con las que estamos trabajando en este estudio, ofrece errores típicos correctos (Recio, 2012).

Tabla 7.16.
Resultados del estudio de la dimensionalidad

	Primaria				**Secundaria**			
	Papel		**Online**		**Papel**		**Online**	
	DWLS	DWLS Robusto	DWLS	DWLS Robusto	DWLS	DWLS Robusto	DWLS	DWLS Robusto
% Varianza Explicada	23%		23%		27%		26%	
RMSEA	0,026	0,028	0,017	0,022	0,018	0,021	0,014	0,018
CFI	0,964	0,937	0,986	0,962	0,986	0,965	0,992	0,976
TLI	0,961	0,933	0,985	0,959	0,985	0,963	0,992	0,975

Fuente: Elaboración propia
Nota: Chi cuadrado (p<0,000)

Los resultados confirman la equivalencia en ambas versiones a consecuencia del estudio dimensional realizado. En el caso de Primaria, es evidente la equivalencia, dado que el porcentaje de varianza explicada tras la extracción de un factor es del 23%, tanto en la versión en papel como en la versión online. En Secundaria, también se confirma la similitud esperable, el porcentaje de varianza explicada tras la extracción de un factor es de un 27% en la versión en papel y un 26% en la versión online.

Siguiendo las ideas de Abad, Olea, Ponsoda y García (2011), se han estimado los tres índices de ajuste más relevantes (RMSEA, CFI y TLI). En Primaria y Secundaria, los ajustes alcanzados son muy adecuados. Si nos detenemos en el índice RMSEA, obtenemos un ajuste muy adecuado en ambas versiones, siempre alcanzando valores inferiores a 0,08.

Lo mismo sucede con los índices CFI y TLI, donde se han de alcanzar valores por encima de 0,95. En el caso de Primaria, según el estimador DWLS, ambas versiones obtienen un índice de ajuste muy adecuado (por encima de 0,95). El estimador DWLS Robusto ofrece valores adecuados, aunque algo inferiores en la versión en papel (cercano a 0,95). En lo que se refiere a Secundaria, los índices CFI y TLI, según el estimador DWLS y DWLS Robusto, ofrecen en todos los casos un ajuste muy adecuado (superior en todos los casos a 0,95).

Estos resultados aportan evidencias del buen ajuste del modelo unidimensionalidad en ambas versiones de la prueba, tanto en Primaria como Secundaria. De forma más

precisa en el anexo 35 (Primaria) y anexo 36 (Secundaria), se representa gráficamente el modelo. Unido a estos resultados y con la intención de garantizar el supuesto de unidimensionalidad, se presenta el análisis de invarianza factorial atendiendo al modo de aplicación.

Invarianza factorial

Los cuatro modelos que se utilizan habitualmente para analizar la invarianza o equivalencia entre versiones, quedan recogidos en la siguiente tabla:

Tabla 7.17.
Modelos e hipótesis para el análisis de la invarianza factorial

MODELO		HIPÓTESIS CONTRASTADA	SIGNIFICADO CONCEPTUAL DE LA HIPÓTESIS
Modelo 1	Invarianza de configuración *"configural invariance"*	Misma estructura factorial en ambas versiones.	Ambas versiones asocian los mismos subjconjuntos de ítems con los mismos constructos.
Modelo 2	Invarianza métrica débil *"weak invariance"*	Igualdad de cargas factoriales	La fuerza de las relaciones entre cada ítem y su constructo subyacente es la misma en ambas versiones.
Modelo 3	Invarianza escalar *"strong invariance"*	Igualdad de interceptos	Las diferencias entre versiones que indican los ítems son las mismas en todos los ítems.
Modelo 4	Invarianza de las varianzas de error *"equal loadings"*	Igualdad de varianzas error	Los ítems tienen la misma consistencia interna en ambas versiones.
Modelo 1-2	Invarianza de la varianza de los factores	Igualdad de las varianzas de los factores	La variabilidad con respecto a los constructos es la misma en ambas versiones.
Modelo 2-3	Invarianza de la covarianza de los factores	Igualdad de las covarianzas entre los factores	Las relaciones entre los constructos son las mismas en ambas versiones.
Modelo 3-4	Invarianza de las medias latentes	Igualdad de medias	La media de cada constructo es la misma en ambas versiones

Fuente: Adaptación de Cuevas (2013, p.59)

Para el análisis de invarianza factorial, comprobaremos si el valor de los pesos factoriales, así como los interceptos y medias de los factores, no varían significativamente entre las versiones. Para ello, llevaremos a cabo comparaciones entre los cuatro modelos y atenderemos a la significación estadística de la diferencia entre los valores de *chi cuadrado*. También, atendiendo a Cheung y Rensvold (2002), se valorará la diferencia entre los valores alcanzados en el índice *CFI* (comparative fix index), que se permite mantener el modelo cuando se alcanza valores inferiores a 0,01.

A continuación, en la tabla 7.18 para Primaria y 7.19 para Secundaria, se llevan a cabo las comparaciones entre los modelos.

Tabla 7.18.
Análisis de la invarianzafactorial atendiendo al modo de aplicación de la prueba en Primaria

Modelo	**χ^2**	**g.l**	**p**	**CFI**
Modelo 1	3329.502	1.054	0,000	0,956
Modelo 2	3816.694	1.087	0,000	0,948
Modelo 3	4054.718	1.120	0,000	0,944
Modelo 4	4448.021	1.121	0,000	0,936
Comparaciones entre modelos	**$\Delta\chi^2$**	**Δ g.l**	**p**	**Δ CFI**
Modelo 1 - Modelo 2	487.192	33	0,000	0,009
Modelo 2 - Modelo 3	238.024	33	0,000	0,004
Modelo 3 - Modelo 4	393.303	1	0,000	0,008

Fuente: Elaboración propia

Al comparar los modelos en Primaria (ver tabla 7.18), podemos observar cómo en todos los casos el valor de chi cuadrado aumenta significativamente ($p< 0,01$). Aún sí, la diferencia que existe en el CFI, es inferior a 0,01 en todas las comparaciones ($p< 0,01$).

Dado que las comparaciones entre los cuatro modelos muestran diferencias significativas entre los valores de chi cuadrado y entre los valores en el índice CFI, podemos concluir la existencia de igualdad de las varianzas de los factores, igualdad de las covarianzas entre los factores e igualdad de medias en ambas versiones.

A continuación, presentamos los resultados obtenidos en Secundaria.

Tabla 7.19.
Análisis de la invarianza factorial atendiendo al modo de aplicación de la prueba en Secundaria

Modelo	**χ^2**	**g.l**	**p**	**CFI**
Modelo 1	3530.774	990	0,000	0,981
Modelo 2	4375.991	1022	0,000	0,975
Modelo 3	4969.196	1054	0,000	0,970
Modelo 4	5112.239	1055	0,000	0,969
Comparaciones entre modelos	**$\Delta\chi^2$**	**Δ g.l**	**p**	**Δ CFI**
Modelo 1 - Modelo 2	845.217	32	0,000	0,006
Modelo 2 - Modelo 3	593.205	32	0,000	0,004
Modelo 3 - Modelo 4	143.043	1	0,000	0,001

Fuente: Elaboración propia

Los resultados arrojan resultados similares a los de Primaria, dado que las comparaciones entre los cuatro modelos muestran diferencias significativas entre los valores de chi cuadrado y entre los del índice CFI. De nuevo, podemos hablar de igualdad de las varianzas de los factores, igualdad de las covarianzas entre los factores e igualdad de medias en ambas versiones.

Puesta a prueba la tercera hipótesis, la estructura factorial refleja la invarianza entre ambas versiones.

Conocida la estructura factorial de las muestras, se procede en el siguiente apartado, al estudio de la equivalencia entre ambas versiones y al estudio del Funcionamiento Diferencial de Versiones.

7.5. Análisis estadísticos para el estudio de la equivalencia de las versiones.

7.5.1. Igualdad de varianzas y prueba T para muestras independientes

En el siguiente apartado, se recogen los resultados relativos a la prueba de igualdad de varianzas y la prueba T para muestras independientes, aplicada para el contraste de cuarta hipótesis, cuya pretensión es verificar si existe o no asociación entre la puntuación en la prueba y el modo en que ha sido aplicada.

- **Igualdad de varianzas**

La hipótesis a contrastar en este apartado corresponde con, la de igualdad de varianzas en ambos grupos, de referencia y focal, empleando el estadístico de homogeneidad F de Snedecor, detallado en la tabla 7.20.

Tabla 7.20.
Resultados del contraste de hipótesis para la igualdad de varianzas

Hipótesis de contraste:	*Resultado Primaria:*	*Resultado Secundaria:*
H_0: $\sigma^2_{GR} = \sigma^2_{GF}$	$\sigma^2_{GR} / \sigma^2_{GF} =$ 31,49/32,29 = 0,975	$\sigma^2_{GR} / \sigma^2_{GF} =$ 28,14/30,42 = 0,932
Estadístico de Homogeneidad F de Snedecor: $F = \sigma^2_{GR} / \sigma^2_{GF}$	gl_{GR} = n-1 = 5485; gl_{GF} = n-1 = 1078	gl_{GR} = n-1 = 14513; gl_{GF} = n-1 = 2206
Interpretación del estadístico F:	*Interpretación:*	*Interpretación:*
F<1 (aceptamos hipótesis nula, asumimos homogeneidad de varianzas) F > 1 (rechazamos hipótesis nula).	Rechazamos la hipótesis de igualdad de varianzas dado que F>1 (F=1,12). Valor obtenido en la tabla F de Snedecor con un alfa de 0,01 y F=0,975.	Rechazamos la hipótesis de igualdad de varianzas dado que F>1 (F=1,08). Valor obtenido en la tabla F de Snedecor con un alfa de 0,01 y F=0,932.

Fuente: Elaboración propia

Los resultados demuestran la no existencia de homogeneidad de las varianzas, lo que refleja la falta de equivalencia entre ambas versiones. A pesar de rechazar la homogeneidad de varianzas, debemos considerar que los valores de las varianzas en ambos grupos son muy aproximados, y las diferencias encontradas son debidas a la elevada potencia estadística con motivo del elevado tamaño muestral del estudio.

- **Prueba T para muestras independientes**

Para continuar contrastando la equivalencia entre ambas versiones, se lleva a cabo un estudio detallado de las medias entre ambas versiones, empleando la prueba t para muestras independientes (tabla 7.21).

Tabla 7.21.
Resumen de la prueba t para muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% I.C Inferior	95% I.C Superior
Primaria	**Si**	0,436	0,509	9,510	6563	**0,000**	1,781	0,187	1,414	2,148
	No			9,430	1520,18	0,000	1,781	0,187	1,414	2,151
Secundaria	Si	21,840	**0,000**	6,035	16716	0,000	,73535	,12184	,49653	,97416
	No			5,865	2861,81	**0,000**	,73535	,12538	,48950	,98119

Fuente: Elaboración propia
Nota: Si (si se han asumido varianzas iguales).
No (no se han asumido varianzas iguales).

En la prueba de Levene para el estudio de la igualdad de varianzas, observamos cómo la significación asociada al estadístico F en Primaria (p= 0,509) no es significativa, dado que es superior a α=0,01. Los resultados nos llevan a asumir la homogeneidad de varianzas. Atendiendo a la prueba t de Student, asumiendo varianzas iguales, la significación asociada al estadístico T (p=0,000), nos lleva a rechazar la hipótesis de igualdad de medias y a concluir que existe una asociación entre la puntuación en la prueba y el modo en que se ha aplicado, ya que las diferencias entre ambas medias son estadísticamente significativas.

En el caso de Secundaria, rechazamos ambas hipótesis. No se asume ni la homogeneidad de varianzas (p=0,000), ni la igualdad de medias (p=0,000); por lo que de nuevo se observa la existencia de diferencia de medias en función del modo de aplicación.

Los resultados alcanzados no son los esperados, dado que no es posible hablar de equivalencia entre ambas versiones en Secundaria, dado que las medias no son iguales y las varianzas, aunque parecidas, son estadísticamente diferentes. Para evidenciar dicha afirmación de forma más concisa, se realiza un estudio del comportamiento de cada ítem en ambas versiones aplicando la prueba Chi cuadrado.

7.5.2. Prueba Chi Cuadrado

En este apartado se proyectan los resultados obtenidos de la prueba Chi cuadrado de Pearson para el contraste de la quinta hipótesis.

En la tabla 7.22 para Primaria y 7.23 para Secundaria, verificamos si existe o no asociación entre la puntuación en cada ítem y el modo en que se aplica el ítem.

Tabla 7.22.
Resumen de la prueba Chi cuadrado y medidas de asociación en Primaria

	Pruebas de X^2		**Medidas de asociación**					
	X^2 de Pearson	Sig. asintótica (bilateral)	Coef. de contingencia	Sig. aproximada	V de Cramer	Sig. Aproximada	Phi	Sig. Aproximada
Ítem 1	25,953	**0,000**	0,063	**0,000**	0,063	**0,000**	-0,063	**0,000**
Ítem 2	2,605	0,107	0,020	0,107	0,020	0,107	-0,020	0,107
Ítem 3	15,379	**0,000**	0,04	**0,000**	0,048	**0,000**	-0,048	**0,000**
Ítem 4	57,701	**0,000**	0,094	**0,000**	0,094	**0,000**	-0,094	**0,000**
Ítem 5	2,202	0,138	0,018	0,138	0,018	0,138	-0,018	0,138
Ítem 6	15,417	**0,000**	0,048	**0,000**	0,048	**0,000**	-0,048	**0,000**
Ítem 7	30,414	**0,000**	0,068	**0,000**	0,068	**0,000**	-0,068	**0,000**
Ítem 8	16,964	**0,000**	0,051	**0,000**	0,051	**0,000**	-0,051	**0,000**
Ítem 9	19,325	**0,000**	0,054	**0,000**	0,054	**0,000**	-0,054	**0,000**
Ítem 10	22,777	**0,000**	0,059	**0,000**	0,059	**0,000**	-0,059	**0,000**
Ítem 11	32,445	**0,000**	0,070	**0,000**	0,070	**0,000**	-0,070	**0,000**
Ítem 12	11,287	**0,000**	0,041	**0,000**	0,041	**0,000**	-0,041	**0,000**
Ítem 13	40,358	**0,000**	0,078	**0,000**	0,078	**0,000**	-0,078	**0,000**
Ítem 14	25,636	**0,000**	0,062	**0,000**	0,062	**0,000**	-0,062	**0,000**
Ítem 15	0,873	0,350	0,012	0,350	0,012	0,350	-0,012	0,350
Ítem 16	34,466	**0,000**	0,072	**0,000**	0,072	**0,000**	-0,072	**0,000**
Ítem 17	24,246	**0,000**	0,061	**0,000**	0,061	**0,000**	-0,061	**0,000**
Ítem 18	18,213	**0,000**	0,053	**0,000**	0,053	**0,000**	-0,053	**0,000**
Ítem 19	20,773	**0,000**	0,056	**0,000**	0,056	**0,000**	-0,056	**0,000**
Ítem 20	48,337	**0,000**	0,086	**0,000**	0,086	**0,000**	-0,086	**0,000**
Ítem 21	0,729	0,393	0,011	0,393	0,011	0,393	0,011	0,393
Ítem 22	24,729	**0,000**	0,061	**0,000**	0,061	**0,000**	-0,061	**0,000**
Ítem 23	0,009	0,925	0,001	0,925	0,001	0,925	0,001	0,925
Ítem 24	7,821	**0,005**	0,035	**0,005**	0,035	**0,005**	0,035	**0,005**
Ítem 25	15,601	**0,000**	0,049	**0,000**	0,049	**0,000**	-0,049	**0,000**
Ítem 26	18,584	**0,000**	0,053	**0,000**	0,053	**0,000**	-0,053	**0,000**
Ítem 27	14,219	**0,000**	0,047	**0,000**	0,047	**0,000**	-0,047	**0,000**
Ítem 28	33,959	**0,000**	0,072	**0,000**	0,072	**0,000**	-0,072	**0,000**
Ítem 29	31,588	**0,000**	0,069	**0,000**	0,069	**0,000**	-0,069	**0,000**
Ítem 30	21,436	**0,000**	0,057	**0,000**	0,057	**0,000**	-0,057	**0,000**
Ítem 31	0,145	0,703	0,005	0,703	0,005	0,703	0,005	0,703
Ítem 32	1,603	0,205	0,016	0,205	0,016	0,205	-0,016	0,205
Ítem 33	0,794	0,373	0,011	0,373	0,011	0,373	-0,011	0,373
Ítem 34	6,156	0,013	0,031	0,013	0,031	0,013	0,031	0,013

Fuente: Elaboración propia
Nota: En negrita se señalan los ítems cuya p< 0,01 con gl= 1

En la tabla 7.22, se revelan los resultados para cada uno de los ítems. Se muestran en negrita aquellos ítems que son estadísticamente significativos (concretamente 25 ítems de 34). En esos 25 casos, se concluye que existe asociación significativa entre la puntuación en cada uno de estos ítems y el modo en que han sido aplicados.

Además del estadístico Chi-cuadrado, se especifican las medidas de asociación (coeficiente de contingencia, V de Cramer y valor Phi) y su nivel crítico, que se caracterizan por corregir el valor del estadístico Chi-cuadrado para minimizar el efecto del tamaño de la muestra.

Los coeficientes de contingencia, V de Cramer y el valor de Phi, presentan valores muy pequeños (todos inferiores a 0,01), lo que evidencia que tras minimizar el efecto del tamaño de la muestra, se observan pequeñas diferencias entre las medias de los ítems en papel y las medias de los ítems online.

Si observamos la tabla 7.23, encontramos los ítems de la prueba en Secundaria.

Tabla 7.23
Resumen de la prueba Chi cuadrado y medidas de asociación en Secundaria

	Pruebas de X^2		**Medidas de asociación**					
	X^2 de Pearson	Sig. asintótica (bilateral)	Coef. de contingencia	Sig. aproximada	V de Cramer	Sig. Aproximada	Phi	Sig. Aproximada
Ítem 1	9,165	**0,002**	0,023	**0,002**	0,023	**0,002**	0,023	**0,002**
Ítem 2	1,859	0,173	0,011	0,173	0,011	0,173	0,011	0,173
Ítem 3	16,894	**0,000**	-0,032	**0,000**	0,032	**0,000**	0,032	**0,000**
Ítem 4	5,578	0,018	-0,018	0,018	0,018	0,018	0,018	0,018
Ítem 5	11,157	**0,001**	0,026	**0,001**	0,026	**0,001**	0,026	**0,001**
Ítem 6	0,629	0,428	0,006	0,428	0,006	0,428	0,006	0,428
Ítem 7	0,201	0,654	-0,003	0,654	0,003	0,654	0,003	0,654
Ítem 8	3,780	0,052	0,015	0,052	0,015	0,052	0,015	0,052
Ítem 9	20,130	**0,000**	-0,035	**0,000**	0,035	**0,000**	0,035	**0,000**
Ítem 10	1,438	0,230	-0,009	0,230	0,009	0,230	0,009	0,230
Ítem 11	19,351	**0,000**	-0,034	**0,000**	0,034	**0,000**	0,034	**0,000**
Ítem 12	151,487	**0,000**	-0,095	**0,000**	0,095	**0,000**	0,095	**0,000**
Ítem 13	20,447	**0,000**	-0,035	**0,000**	0,035	**0,000**	0,035	**0,000**
Ítem 14	79,765	**0,000**	-0,069	**0,000**	0,069	**0,000**	0,069	**0,000**
Ítem 15	73,090	**0,000**	-0,066	**0,000**	0,066	**0,000**	0,066	**0,000**
Ítem 16	15,025	**0,000**	0,030	**0,000**	0,030	**0,000**	0,030	**0,000**
Ítem 17	1,691	0,193	0,010	-0,193	0,061	-0,193	0,061	-0,193
Ítem 18	52,688	**0,000**	-0,056	**0,000**	0,056	**0,000**	0,056	**0,000**
Ítem 19	5,574	0,018	-0,018	0,018	0,018	0,018	0,018	0,018
Ítem 20	58,811	**0,000**	0,059	**0,000**	0,059	**0,000**	0,059	**0,000**
Ítem 21	83,246	**0,000**	-0,071	**0,000**	0,071	**0,000**	0,071	**0,000**
Ítem 22	45,000	**0,000**	-0,052	**0,000**	0,052	**0,000**	0,052	**0,000**
Ítem 23	8,380	**0,004**	-0,022	**0,004**	0,022	**0,004**	0,022	**0,004**
Ítem 24	17,121	**0,000**	0,032	**0,000**	0,032	**0,000**	0,032	**0,000**
Ítem 25	17,932	**0,000**	-0,033	**0,000**	0,033	**0,000**	0,033	**0,000**
Ítem 26	41,488	**0,000**	-0,050	**0,000**	0,050	**0,000**	0,050	**0,000**
Ítem 27	10,621	**0,001**	-0,025	**0,001**	0,025	**0,001**	0,025	**0,001**
Ítem 28	0,063	0,801	-0,002	0,801	0,002	0,801	0,002	0,801
Ítem 29	5,687	0,017	-0,018	0,017	0,018	0,017	0,018	0,017
Ítem 30	3,315	0,069	0,014	0,069	0,014	0,069	0,014	0,069
Ítem 31	7,838	**0,005**	0,022	**0,005**	0,022	**0,005**	0,022	**0,005**
Ítem 32	0,275	0,600	0,004	0,600	0,004	0,600	0,004	0,600
Ítem 33	11,785	**0,001**	-0,027	**0,001**	0,027	**0,001**	0,027	**0,001**

Fuente: Elaboración propia
Nota: En negrita se señalan los ítems cuya p< 0,01 con gl= 1

De nuevo, en negrita se indican los ítems que presentan diferencias estadísticamente significativas (p inferior a α=0,01). Se detectan 21 ítems de los 33 que conforman la prueba, con diferencias significativas entre sus medias en ambas versiones; pero atendiendo a las medidas de asociación, sucede al igual que en Primaria, que el tamaño del efecto es muy pequeño, lo que nos lleva a concluir que las diferencias observadas, tras minimizar el efecto del tamaño de la muestra, son débiles.

7.6. Estudio del Funcionamiento Diferencial de Versiones

En el siguiente apartado, se aborda la sexta y última hipótesis planteada en la tesis doctoral. Teniendo en cuenta el modo en el que se ha aplicado el test (papel y online), no existe Funcionamiento Diferencial de Versiones, por tanto ambas versiones de la prueba de rendimiento en comprensión lectora son equivalentes.

7.6.1. Corrección Benjamini y Hochberg: *False Discovery Rate*

Cuando contrastamos cualquier test de hipótesis, es posible que se cometan dos tipos de errores; estos errores son comúnmente conocidos como Error Tipo I (falso positivo), que indica el rechazo de una hipótesis nula cuando realmente dicha hipótesis es cierta y debería ser aceptada; y el Error Tipo II (falso negativo), corresponde a lo contrario, aceptar una hipótesis nula cuando debería haber sido rechazada (Ferrero, 2011). Resumimos esta idea en la tabla 7.24.

Tabla 7.24.
Errores Tipo I y Tipo II

	Acepto H_0	Rechazo H_0
H_0 Verdadera	Decisión adecuada (robustez)	Decisión incorrecta *Error Tipo I*
H_0 Falsa	Decisión incorrecta *Error Tipo II*	Decisión adecuada (potencia)

Fuente: Elaboración propia

Es muy común llevar a cabo múltiples comparaciones, como sucede en el estudio realizado en esta tesis doctoral. La situación con la que podemos encontrarnos ante *m* hipótesis es la recogida en la tabla 7.25.

Tabla 7.25.
Posibles resultados ante m hipótesis

	Acepto H_0	Rechazo H_0	Total
H_0 Verdadera	U (robustez)	V *Error Tipo I*	m_0
H_0 Falsa	T *Error Tipo II*	S (potencia)	m - m0 = m1
Total	m - R	R	m

Fuente: Elaboración propia

Los valores de *m* y *R* son conocidos por el investigador, concretamente se refieren a:

m: número total de hipótesis nulas del estudio

R: número de hipótesis rechazadas

Pero existen variables no observables, es decir desconocidas por el investigador, que son:

U: número de hipótesis nulas verdaderas aceptadas

V: número de hipótesis verdaderas rechazadas – Error Tipo I

S: número de hipótesis nulas falsas rechazadas

T: número de hipótesis falsas declaradas verdaderas – Error Tipo II

En estos casos debe ser controlado el Error Tipo I principalmente, dado que ante los diferentes contrastes de hipótesis, tiende a incrementarse la posibilidad de falsos positivos; es decir, debemos controlar que no rechacemos la hipótesis nula (no existe Funcionamiento Diferencial de Versiones), cuando realmente existe Funcionamiento Diferencial de Versiones.

Ante esta situación, y con la intención de reducir la proporción de falsos positivos, se procederá al cálculo de los niveles de significación corregidos, atendiendo al método "*False Discovery Rate*" (FDR) descrito por Benjamini y Hochberg (1995).

El método "*False Discovery Rate*" o Tasa de Falsos Descubrimientos, puede definirse como la proporción esperada de Errores Tipo I entre las hipótesis rechazadas. Atendiendo a Salazar (2011, p.11), para este tipo de error *"debe tenerse en cuenta que la proporción de errores de tipo I entre las hipótesis rechazadas, V/R, es cero cuando el número de hipótesis que se rechazan, R, es cero"*.

Dicha afirmación, según Benjamini y Hochberg (1995), matemáticamente es formulada como:

$$Q = V/ (V + S)$$
$$Q = 0 \text{ cuando } V + S = 0$$
$$Q_e = E(Q) = E\{V/(V+S)\} = E(V/R).$$

El proceso de Benjamini y Hochberg (1995) puede ser controlado, para lo cual es necesario:

1. Probar las *m* hipótesis nulas del estudio ($H_1, H_2, H_3 \ldots H_m$):

   La hipótesis nula que se contrasta es la no existencia de Funcionamiento Diferencial de Versiones en cada una de las pruebas realizadas.

2. Obtener los valores $p_1, p_2, p_3 \ldots p_m$ calculados por múltiples pruebas en los mismos datos.

   Concretamente los p-valores obtenidos en la Regresión Logística (prueba $\chi 2$ para la diferencia en los modelos 1 y 3).

   Si esta diferencia es significativa podemos hablar de la existencia de DVF.

3. Ordenar los *p* valores resultantes de las *m* pruebas realizadas: $p_{(1)} > p_{(2)} > p_{(3)} \ldots > p_{(m)}$.

4. Estimar el p- valor ajustado:

$P_{(i)} = (m_0/ m)*\alpha$

Donde:

m_0: número de la hipótesis nula a estudiar

m: número total de hipótesis nulas del estudio. En nuestro estudio el número de hipótesis nulas en Primaria es de 34 y en Secundaria de 33.

Así como el nivel de significación (α) fijado para el estudio, que corresponde a α=0,01.

5. Rechazaremos la hipótesis en cada uno de los *m* casos, cuando p-valor sin ajustar es ≤ que el p valor ajustado ($P_{(i)}$).

Los resultados obtenidos tras la corrección de Benjamini y Hochberg (1995), utilizando el método "*False Discovery Rate*" o Tasa de Falsos Descubrimientos, se presentan en la tabla 7.26.

Tabla 7.26.
Resultados Regresión Logística para la detección del Funcionamiento Diferencial de Versiones tras la corrección de Benjamini y Hochberg (1995). Método "False Discovery Rate" o Tasa de Falsos Descubrimientos.

Primaria			Secundaria		
Hipótesis	p-valor sin ajustar	p-valor ajustado $(m_0/ m)*\alpha$	Hipótesis	p-valor sin ajustar	p-valor ajustado $(m_0/ m)*\alpha$
1	0,206	0,0003	1	**0,001**	**0,0003**
2	0,262	0,0006	2	**0,022**	**0,0006**
3	0,869	0,0009	3	**0,003**	**0,0009**
4	**0,000**	**0,0012**	4	0,327	0,0012
5	0,386	0,0015	5	**0,000**	**0,0015**
6	0,736	0,0018	6	0,069	0,0018
7	**0,002**	**0,0021**	7	0,465	0,0021
8	0,899	0,0024	8	**0,006**	0,0024
9	0,330	0,0026	9	**0,000**	**0,0027**
10	0,158	0,0029	10	0,841	0,0030
11	**0,000**	**0,0032**	11	**0,000**	**0,0033**
12	0,160	0,0035	12	**0,000**	**0,0036**
13	**0,000**	**0,0038**	13	**0,000**	**0,0039**
14	0,257	0,0041	14	**0,000**	**0,0042**
15	**0,002**	**0,0044**	15	**0,000**	**0,0045**
16	0,067	0,0047	16	**0,000**	**0,0048**
17	0,281	0,0050	17	0,590	0,0052
18	0,299	0,0053	18	**0,000**	**0,0055**
19	**0,003**	**0,0056**	19	0,124	0,0058
20	**0,000**	**0,0059**	20	**0,000**	**0,0061**
21	**0,007**	**0,0062**	21	**0,000**	**0,0064**
22	0,802	0,0065	22	**0,000**	**0,0067**
23	**0,000**	**0,0068**	23	0,489	0,0070
24	**0,000**	**0,0071**	24	**0,000**	**0,0073**
25	0,118	0,0074	25	**0,003**	**0,0076**
26	0,437	0,0076	26	**0,000**	**0,0079**
27	0,303	0,0079	27	0,257	0,0082
28	0,917	0,0082	28	0,027	0,0085
29	0,460	0,0085	29	0,675	0,0088
30	0,288	0,0088	30	**0,000**	**0,0091**
31	**0,000**	**0,0091**	31	0,215	0,0094
32	0,165	0,0094	32	0,021	0,0097
33	**0,000**	**0,0097**	33	**0,003**	**0,0100**
34	**0,000**	**0,0100**	-	-	-

Fuente: Elaboración propia
Nota: La corrección Benjamini y Hochberg, atiende a un α= 0,01

Tras la corrección llevada a cabo, las decisiones siguen siendo las mismas, pues cuando se detectaba que un ítem presentada DVF con α= 0,01, también es detectado como ítem con DVF con el p-valor ajustado, a excepción del ítem 8 en Secundaria, que presenta DVF cuando se atiende al p- valor no ajustado, pero al realizar la corrección Benjamini y Hochberg este ítem no presenta DVF [p(0,006) > p-ajustado (0,0024)].

7.6.2. Regresión Logística

En el siguiente apartado, se muestran los resultados alcanzados en Primaria y en Secundaria. Es interesante en este punto, recordar los modelos que van a ser utilizar para el análisis con el procedimiento de la Regresión Logística.

Modelo 1: Habilidad o Puntuación Total.

- Únicamente valora el parámetro de la variable habilidad o puntuación total.

Modelo 2: Habilidad o Puntuación total + Grupo.

- Incluye al modelo 1, el parámetro de pertenencia al grupo. Se estudia el DVF uniforme, cuando la diferencia entre el modelo 1 y el modelo 2 es significativa.

Modelo 3: Habilidad o Puntuación total + Grupo + Interacción entre habilidad o Puntuación total y Grupo.

- Incluye al modelo 2, el término de interacción entre la habilidad o puntuación total y el grupo. Cuando se compara el modelo 2 y 3 se estudia el DVF no uniforme.

Atendiendo a dichos modelos, los resultados de Primaria y Secundaria son resumidos en la tabla 7.27 y 7.28 respectivamente.

Tabla 7.27.
Regresión Logística y Funcionamiento Diferencial de Versiones de Primaria

Ítem	chi12	chi13	chi23	p-valor ajustado	Tipo de DVF	Diferencia R^2 Modelo 1 – 2	Diferencia R^2 Modelo 1 – 3	Diferencia R^2 Modelo 2 – 3	Efecto DVF
1	0,076	0,206	0,940	0,0003	No DVF	-	-	-	-
2	0,194	0,262	0,320	0,0006	No DVF	-	-	-	-
3	0,618	0,869	0,860	0,0009	No DVF	-	-	-	-
4	**0,000**	**0,000**	0,726	**0,0012**	DVF uniforme	0,0024	0,0024	0,0000	débil
5	0,705	0,386	0,185	0,0015	No DVF	-	-	-	-
6	0,839	0,736	0,449	0,0018	No DVF	-	-	-	-
7	0,017	**0,002**	**0,007**	**0,0021**	DVF no uniforme	0,0010	0,0024	0,0013	débil
8	0,733	0,899	0,756	0,0024	No DVF	-	-	-	-
9	0,245	0,330	0,353	0,0026	No DVF	-	-	-	-
10	0,058	0,158	0,757	0,0029	No DVF	-	-	-	-
11	0,000	**0,000**	**0,000**	**0,0032**	DVF no uniforme	0,0021	0,0042	0,0021	débil
12	0,966	0,160	0,056	0,0035	No DVF	-	-	-	-
13	0,001	**0,000**	**0,000**	**0,0038**	DVF no uniforme	0,0011	0,0039	0,0027	débil
14	0,116	0,257	0,625	0,0041	No DVF	-	-	-	-
15	**0,001**	**0,002**	0,840	**0,0044**	DVF uniforme	0,0020	0,002	0,0000	débil
16	0,110	0,067	0,090	0,0047	No DVF	-	-	-	-
17	0,497	0,281	0,150	0,0050	No DVF	-	-	-	-
18	0,702	0,299	0,132	0,0053	No DVF	-	-	-	-
19	0,015	**0,003**	**0,017**	**0,0056**	DVF no uniforme	0,0128	0,0013	0,0006	débil
20	**0,000**	**0,000**	0,016	**0,0059**	DVF uniforme	0,0018	0,0024	0,0007	débil
21	**0,002**	**0,007**	0,983	**0,0062**	DVF uniforme	0,0011	0,0011	0,0000	débil
22	0,873	0,802	0,519	0,0065	No DVF	-	-	-	-
23	0,000	**0,000**	**0,000**	0,0068	DVF no uniforme	0,0052	0,011	0,0058	débil
24	**0,000**	**0,000**	0,430	**0,0071**	DVF uniforme	0,0041	0,0042	0,0001	débil
25	0,454	0,118	0,054	0,0074	No DVF	-	-	-	-
26	0,589	0,437	0,243	0,0076	No DVF	-	-	-	-
27	0,175	0,303	0,460	0,0079	No DVF	-	-	-	-
28	0,828	0,917	0,722	0,0082	No DVF	-	-	-	-
29	0,291	0,46	0,508	0,0085	No DVF	-	-	-	-
30	0,711	0,288	0,125	0,0088	No DVF	-	-	-	-
31	**0,000**	**0,000**	0,098	**0,0091**	DVF uniforme	0,0040	0,0045	0,0005	débil
32	0,122	0,165	0,271	0,0094	No DVF	-	-	-	-
33	**0,000**	**0,000**	0,187	**0,0097**	DVF uniforme	0,0015	0,0017	0,0002	débil
34	0,000	**0,000**	**0,004**	**0,0100**	DVF no uniforme	0,0059	0,0072	0,0013	débil

Fuente: Elaboración propia
Nota: En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 3 significativas (atendiendo al p- valor ajustado). Lo que significa la presencia de DVF.
En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 2 significativa (p- valor ajustado) (presenta DVF uniforme)
En negrita se indican las diferencias significativas entre el Modelo 2 - Modelo 3 significativa (p- valor ajustado) (presenta DVF no uniforme)
Diferencia R^2_{12} , R^2_{13} , R^2_{23} ($p<0,035$) (DVF débil o insignificante)

En la tabla 7.27, se puede apreciar que 13 ítems (de los 34) presentan Funcionamiento Diferencial de Versiones (DVF). Dicha evidencia emerge de la prueba $\chi 2$ para la diferencia en los modelos 1 y 3, cuando la diferencia es significativa ($p \leq$ p-ajustado), refleja presencia de DVF.

Cuando la diferencia entre el Modelo 1 y 2 es significativa ($p \leq$ p-ajustado), estamos ante la presencia de Funcionamiento Diferencial de Versiones uniforme. Esto es lo que sucede con los ítems 4, 15, 20, 21, 24, 31 y 33. En cambio, si la diferencia entre el Modelo 2 y 3 es significativa ($p \leq$ p-ajustado), denota presencia de Funcionamiento Diferencial de Versiones no uniforme; como sucede con los ítems 7, 11, 13, 19, 23, 34.

El siguiente valor a interpretar es el correspondiente a la diferencia entre los Pseudo R^2 (medida del efecto del DVF basada en la medida de mínimos cuadrados ponderados) en las comparaciones llevadas a cabo entre los modelos: R^2_{12}, R^2_{13}, $R^2_{23.}$ Este valor nos informa de la intensidad del DVF, siendo:

$R^2 < 0{,}035$: DVF débil o insignificante
$0{,}035 < R^2 < 0{,}07$: DVF moderado
$R^2 > 0{,}07$: DVF relevante o elevado

Dicho lo anterior, a pesar de que estos ítems presenten Funcionamiento Diferencial de Versiones, las diferencias entre los Pseudo R^2 son inferiores a 0,035 en todos los ítems; por lo que la medida del efecto nos permite confirmar que los 13 ítems presentan DVF muy débil o insignificante.

Tabla 7.28.
Regresión Logística y Funcionamiento Diferencial de Versiones de Secundaria

Ítem	chi12	chi13	chi23	p-valor ajustado	Tipo de DVF	Diferencia R^2 Modelo 1 – 2	Diferencia R^2 Modelo 1 – 3	Diferencia R^2 Modelo 2 – 3	Efecto DVF
1	**0,000**	**0,001**	0,414	**0,0003**	DVF uniforme	0,0006	0,0007	0,0000	débil
2	**0,006**	**0,022**	0,764	**0,0006**	DVF uniforme	0,0006	0,0006	0,0000	Débil
3	**0,001**	**0,003**	0,293	**0,0009**	DVF uniforme	0,0007	0,0007	0,0001	Débil
4	0,245	0,327	0,346	0,0012	No DVF	-	-	-	-
5	**0,000**	**0,000**	0,050	**0,0015**	DVF uniforme	0,0022	0,0025	0,0003	Débil
6	0,169	0,069	0,063	0,0018	No DVF	-	-	-	-
7	0,557	0,465	0,276	0,0021	No DVF	-	-	-	-
8	0,002	0,006	0,400	0,0024	No DVF	-	-	-	-
9	**0,000**	**0,000**	0,261	**0,0027**	DVF uniforme	0,0007	0,0008	0,0001	débil
10	0,745	0,841	0,624	0,0030	No DVF	-	-	-	-
11	0,001	**0,000**	**0,001**	**0,0033**	DVF no uniforme	0,0009	0,0016	0,0008	débil
12	0,000	**0,000**	**0,000**	**0,0036**	DVF no uniforme	0,0074	0,0086	0,0012	débil
13	0,003	**0,000**	**0,009**	**0,0039**	DVF no uniforme	0,0007	0,0012	0,0005	débil
14	**0,000**	**0,000**	0,736	**0,0042**	DVF uniforme	0,003	0,0031	0,0000	débil
15	0,000	**0,000**	**0,000**	**0,0045**	DVF no uniforme	0,0027	0,0034	0,0007	débil
16	**0,000**	**0,000**	0,618	**0,0048**	DVF uniforme	0,0013	0,0013	0,0000	débil
17	0,987	0,590	0,305	0,0052	No DVF	-	-	-	-
18	**0,000**	**0,000**	0,201	**0,0055**	DVF uniforme	0,0021	0,0022	0,0001	débil
19	0,053	0,124	0,507	0,0058	No DVF	-	-	-	-
20	**0,000**	**0,000**	0,121	**0,0061**	DVF uniforme	0,0056	0,0057	0,0001	débil
21	0,000	**0,000**	**0,001**	**0,0064**	DVF no uniforme	0,0042	0,005	0,0007	débil
22	**0,000**	**0,000**	0,373	**0,0067**	DVF uniforme	0,0015	0,0016	0,0000	débil
23	0,359	0,489	0,442	0,0070	No DVF	-	-	-	-
24	**0,000**	**0,000**	0,417	**0,0073**	DVF uniforme	0,003	0,003	0,0001	débil
25	**0,001**	**0,003**	0,299	**0,0076**	DVF uniforme	0,0005	0,0005	0,0001	débil
26	**0,000**	**0,000**	0,015	**0,0079**	DVF uniforme	0,0016	0,002	0,0004	débil
27	0,240	0,257	0,247	0,0082	No DVF	-	-	-	-
28	0,044	0,027	0,074	0,0085	No DVF	-	-	-	-
29	0,400	0,675	0,781	0,0088	No DVF	-	-	-	-
30	**0,000**	**0,000**	0,058	**0,0091**	DVF uniforme	0,0018	0,0021	0,0004	débil
31	0,309	0,215	0,153	0,0094	No DVF	-	-	-	-
32	0,033	0,021	0,073	0,0097	No DVF	-	-	-	-
33	0,019	**0,003**	**0,014**	**0,0100**	DVF no uniforme	0,0003	0,0005	0,0003	débil

Fuente: Elaboración propia
Nota: En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 3 significativas (atendiendo al p- valor ajustado). Lo que significa la presencia de DVF.
En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 2 significativa (p- valor ajustado) (presenta DVF uniforme)
En negrita se indican las diferencias significativas entre el Modelo 2 - Modelo 3 significativa (p- valor ajustado) (presenta DVF no uniforme)
Diferencia R^2_{12} , R^2_{13} , R^2_{23} (p<0,035) (DVF débil o insignificante)

Los resultados presentados en la tabla 7.28, correspondiente a Secundaria, muestra las mismas conclusiones que en Primaria. En este caso estamos ante 20 ítems de los 33, que muestran Funcionamiento Diferencial de Versiones debido a las diferencias significativas alcanzadas entre el modelo 1 y 3 ($p \leq$ p-ajustado).

De entre los 20 ítems, detectamos que el Funcionamiento Diferencia de Versiones es uniforme en los ítems 1, 2, 3, 5, 9, 14, 16, 18, 20, 22, 24, 25, 26 y 30; dado que la diferencia es significativa entre el Modelo 1 y 2 ($p \leq$ p-ajustado). Los ítems restantes (11, 12, 13, 15, 21 y 33) presentan Funcionamiento Diferencial de Versiones no uniforme, ya que la diferencia entre el Modelo 2 y 3 es significativa ($p \leq$ p-ajustado).

Al igual que sucedía en Primaria, las diferencia entre los Pseudo R^2 en cada comparación entre los modelos es inferior en todos los casos a 0,035. La medida del efecto verifica que los 20 ítems presentan DVF irrelevante o muy débil.

Análisis global de la prueba

El paquete Lordif realiza un análisis global que permite, por medio de gráficos, visualizar los resultados alcanzados en la prueba, atendiendo tanto al grupo de referencia como al grupo focal.

Las primeras figuras que presentamos (figura 13- Primaria y figura 14 – Secundaria), muestran un histograma de los niveles de habilidad (theta), atendiendo a los sujetos que realizan la prueba en papel (línea continua) y los sujetos que realizan la prueba online (línea discontinua).

En la figura 13, correspondiente a Primaria, visualmente se aprecia el solapamiento que existe entre ambas distribuciones; aunque los sujetos que realizan la prueba en papel tienen niveles de habilidad superiores a los sujetos que realizan la prueba online.

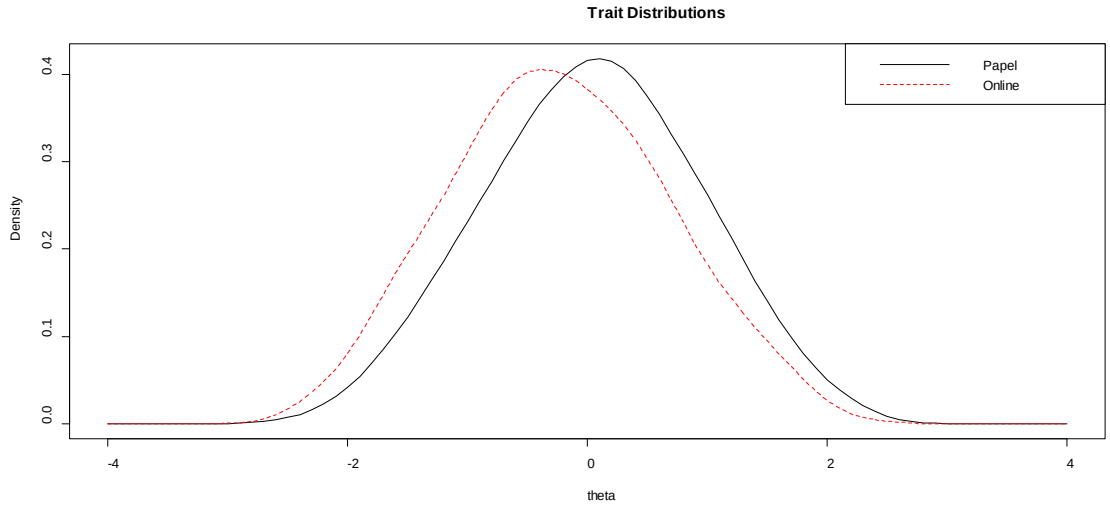


Figura 13. Distribución del rango latente en función del modo de aplicación (papel vs online) en Primaria.

En Secundaria sucede lo mismo, aunque existe algo más de solapamiento entre ambas distribuciones (ver figura 14).

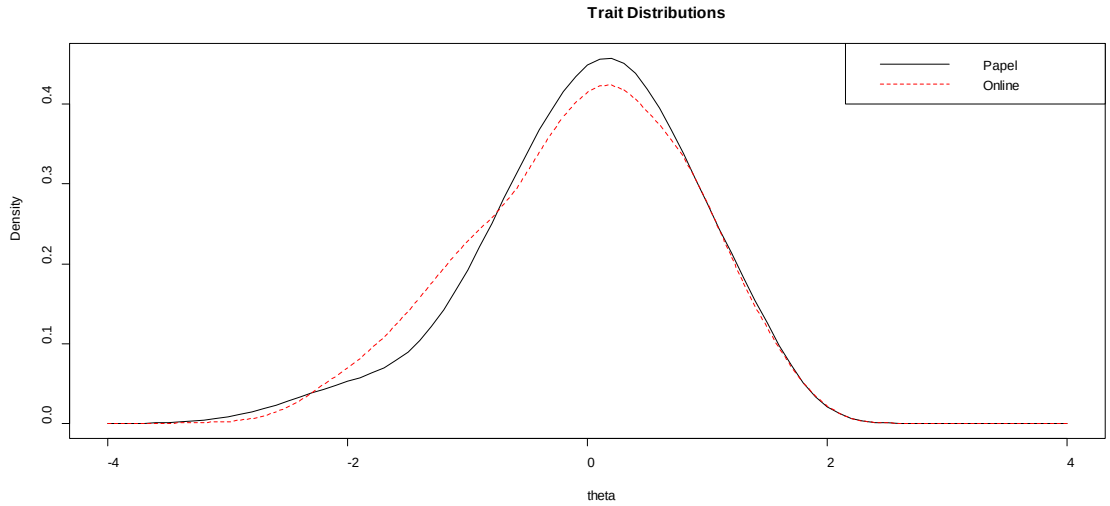


Figura 14. Distribución del rango latente en función del modo de aplicación (papel vs online) en Secundaria.

A continuación, nos centramos en la Curva Característica del Test, y el impacto de los ítems con DVF. En la figura 15 (Primaria) y 16 (Secundaria), podemos observar las Curvas Características de los Test para los sujetos que realizan la prueba en papel y online. El gráfico de la izquierda, muestra las curvas de todos los ítems (tanto los ítems

con DVF como los ítems sin DVF), mientras que el gráfico de la derecha, muestra las curvas para los ítems que presentan DVF.

Estas curvas indican la semejanza alcanzada en la puntuación total esperada en cualquier nivel de habilidad entre los sujetos que realizan la prueba en papel y los que la realizan online.

En la prueba de Primaria (ver figura 15), podemos observar cómo se mantiene la misma tendencia en todos los ítems. Las diferencias apreciables, principalmente en niveles de habilidad alta, se deben a los ítems que presentan DVF; concretamente esta diferencia favorece a los sujetos que realizan la prueba en papel.

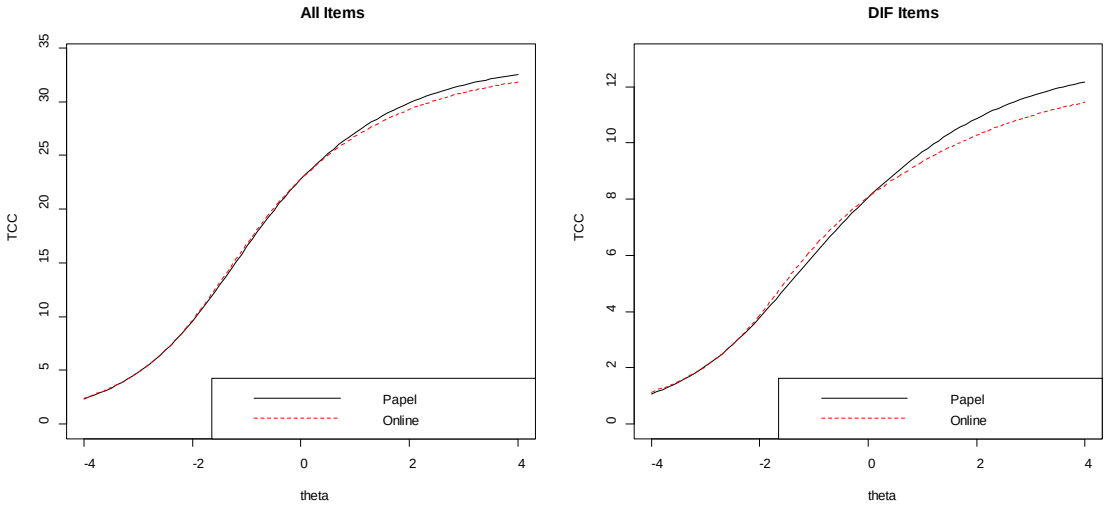


Figura 15: Impacto de los ítems con DVF en la Curva característica del test (Primaria).

En Secundaria, (ver figura 16) observamos que la curva en la que se recoge la información de todos los ítems, de nuevo sigue una misma tendencia, prácticamente exacta, pero cuando atendemos a la curva con los ítems que presentan DVF, apreciamos algunas diferencias en los niveles de habilidad intermedia, que tienden a favorecer mínimamente a los sujetos que realizan la prueba en papel.

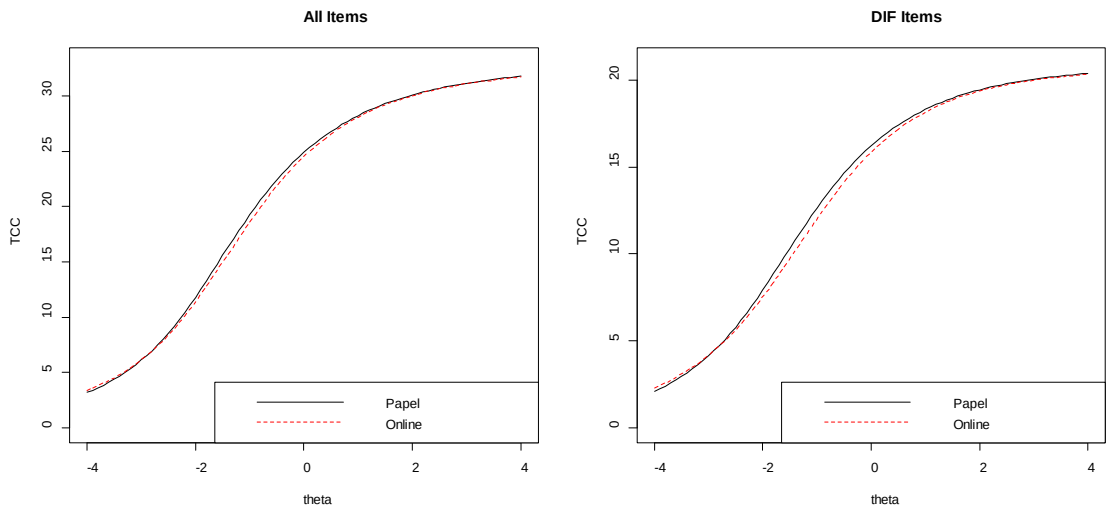


Figura 16: Impacto de los ítems con DVF en la Curva característica del test (Secundaria).

En las siguientes figuras 17 y 18, se representa la diferencia entre la habilidad inicial del sujeto y la habilidad purificada (es decir, ignorando el Funcionamiento Diferencial de Versiones – DVF). Concretamente a la izquierda, observamos un diagrama de cajas, donde se muestran las diferencias entre la puntuación inicial ("theta inicial"), y la purificada (es decir, la puntuación una vez eliminados los ítems con DVF). Y a la derecha, se presentan en un diagrama de dispersión, las diferencias entre la puntuación inicial y la puntuación purificada de forma separada para la aplicación en papel y online. La línea continua nos indica la no diferencia, mientras que la línea de puntos nos indica la media de las diferencias. Podemos apreciar cómo coincide en ambos casos.

En Primaria (ver figura 17), la diferencia en la versión en papel se localiza en torno a la línea continua (no existencia de diferencia), que coincide con la media de las diferencias, que es 0. Este dato nos indica la no existencia de ítems con DVF, puesto que la diferencia entre la habilidad inicial – habilidad purificada (eliminando el DVF) es 0. En cambio, si estudiamos los resultados de la prueba online, vemos que se distribuyen por toda la gráfica; los sujetos cuya habilidad en la prueba online es baja, podemos apreciar cómo la diferencia entre la habilidad inicial – habilidad purificada es positiva, por lo que se tiende a subestimar la habilidad de estos sujetos. Al eliminar el DVF en la prueba, se obtienen habilidades inferiores que las obtenidas en la habilidad inicial.

Sucede lo contrario con los sujetos con habilidades altas, donde se sobreestima la puntuación de estos sujetos, puesto que la diferencia entre la habilidad inicial – habilidad purificada es negativa; por lo que eliminando los ítems con DVF, estos sujetos obtendrían habilidades superiores a la habilidad inicial obtenida.

En vista de los resultados obtenidos, la prueba online puede influir en las habilidades de los sujetos, por ello debemos estudiar en profundidad estos ítems por la posible presencia de Funcionamiento Diferencial de Versiones.

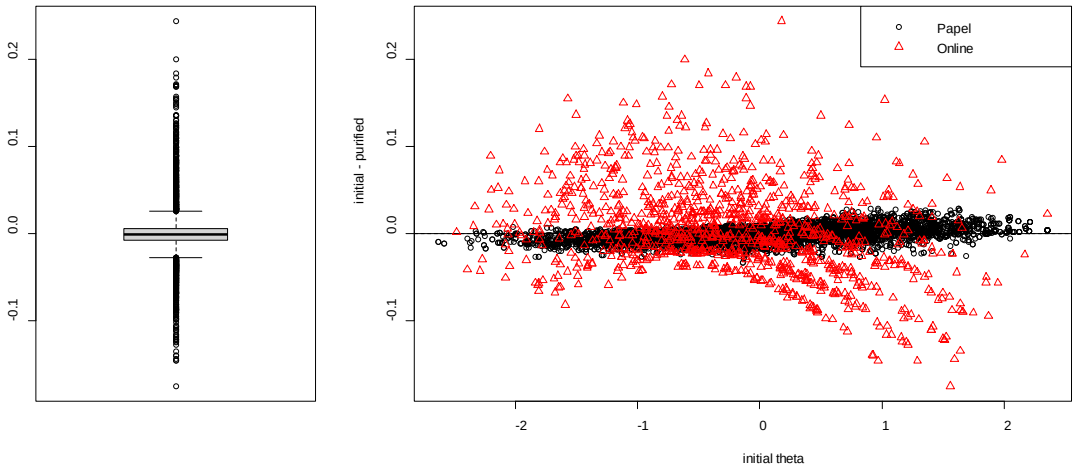


Figura 17. Impacto del DVF de manera individual (Primaria).

En lo que concierne a Secundaria (ver figura 18), la diferencia entre la habilidad inicial y purificada, al igual que sucedía en Primaria, no existe. Los sujetos se agrupan en torno a 0, lo que nos lleva a pensar que no existe DVF.

La prueba online presenta resultados muy diferentes a los alcanzados en la prueba en papel. En muchos casos, la diferencia entre las habilidades iniciales medio-altas y las habilidades purificadas, es negativa; por lo que al igual que en Primaria, a los sujetos con habilidad altas, se les sobreestima su habilidad; lo mismo que sucede con los sujetos con habilidades iniciales más bajas.

De nuevo, es necesario estudiar en profundidad el Funcionamiento Diferencial de Versiones en ambas pruebas, porque la prueba online puede ofrecer puntuaciones diferentes.

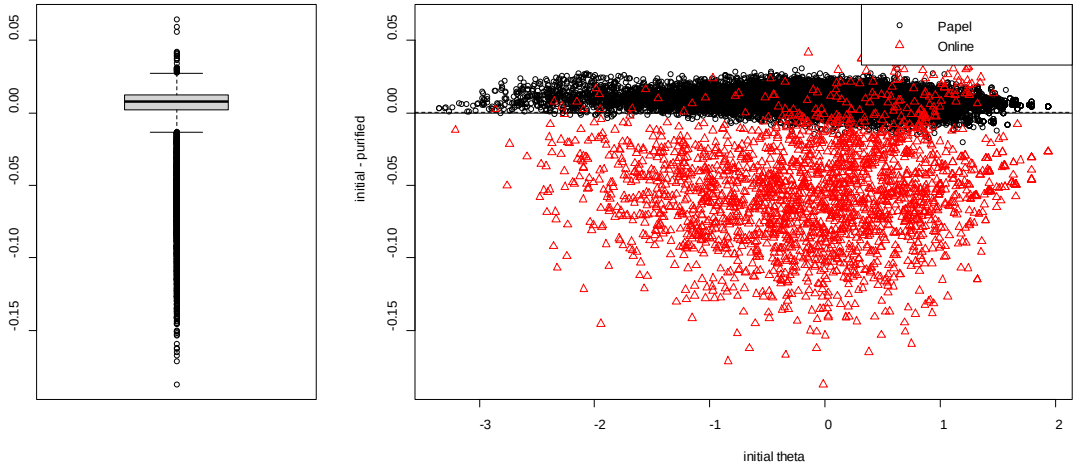


Figura 18. Impacto del DVF de manera individual (Secundaria).

Tras llevar a cabo este resumen, se procede a estudiar detalladamente cada uno de los ítems que presentan DVF.

En primer lugar se mostrarán los ítems con DVF en Primaria, que concretamente son: 4, 7, 11, 13, 15, 19-21, 23, 24, 31, 33 y 34. Y posteriormente, los ítems con DVF en Secundaria: 1-3, 5, 8, 9, 11-16, 18, 20-22, 24-26, 30, 33.

7.6.2.1. Descripción de los ítems con DVF en Primaria

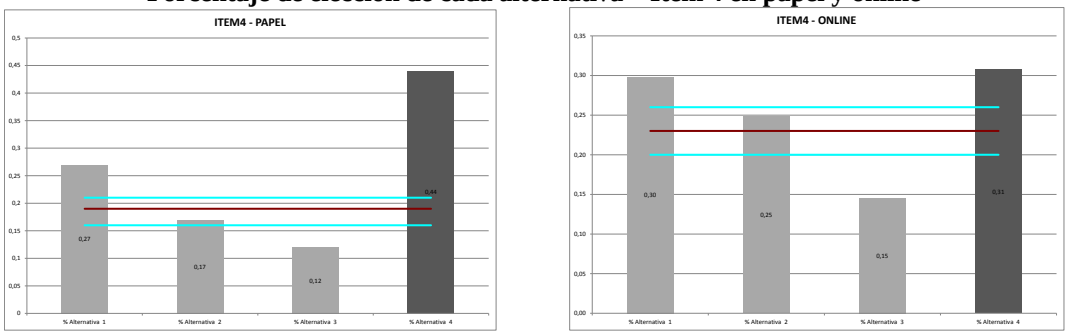
Para cada ítem detectado con DFV, se presenta una tabla con sus características según la TCT y TRI, así como diferentes técnicas de detección del DVF. Con esta información, tenemos un resumen de la situación psciométrica de cada uno de los ítems.

Tabla 7.29.
Características y Funcionamiento Diferencial de Versiones en el Ítem 4

Descripción desde la TCT

Ítem 4	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,43	0,24	0,49	0,36	0,28	4	24,6	20,6	0,43	0,495
Online	0,31	0,21	0,46	0,36	0,29	4	23,5	19,1	0,31	0,462

Porcentaje de elección de cada alternativa – Ítem 4 en papel y online



Modelo TRI de 2 Parámetros

Ítem 4	Parámetro a	Parámetro b	p
Papel	0,706	0,339	0,002
Online	0,789	1,159	0,033

Técnicas detección DVF Ítem 4

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,726	Uniforme	0,0024	0,0024	0,0000	Débil					

*Diferencias significativas (p≤0,0012) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

Como podemos observar en la información resumida sobre las características del ítem 4, éste tiene una media superior en los sujetos que realizan el ítem en papel, presentando un índice de facilidad mayor que el ítem online; a pesar de ello, se trata de un ítem difícil. Las correlaciones Biserial Puntual en ambos casos son prácticamente las mismas (0,28 en papel y 0,29 online). Se trata de un ítem que discrimina poco, tomando valores aproximados en ambos casos a 0,30.

Atendiendo a un modelo TRI de dos parámetros, se arrojan resultados semejantes en ambos grupos, siendo el parámetro de discriminación y de dificultad superior en la prueba online.

En lo que respecta a las técnicas de detección del DVF, además de incluir el procedimiento de Regresión Logística desarrollado anteriormente, se incluyen, para tener una visión global, otros métodos de detección. De esta forma observamos cómo algunos métodos detectan el ítem con DVF y otros no; además del procedimiento Regresión Logística que detecta el ítem 4 con DVF, debemos destacar que otros métodos también lo detectan, como es el caso del método de Rajú, Lord y MH.

A continuación, presentamos algunos gráficos que resumen la información sobre el Funcionamiento Diferencial de Versiones en el ítem 4.

En la figura 19, se muestran las funciones de respuesta al ítem para los dos grupos, en función de las estimaciones de los parámetros (parámetro de discriminación y de dificultad[16]). Se aprecia como ambos parámetros son similares en ambas versiones.

[16] Los parámetros en la prueba en papel pueden variar, puesto que para comprobar el ajuste se ha utilizado una muestra de 1000 sujetos. En las estimaciones del DIF se ha utilizado la muestra completa estimada por medio de Porpensity Score.

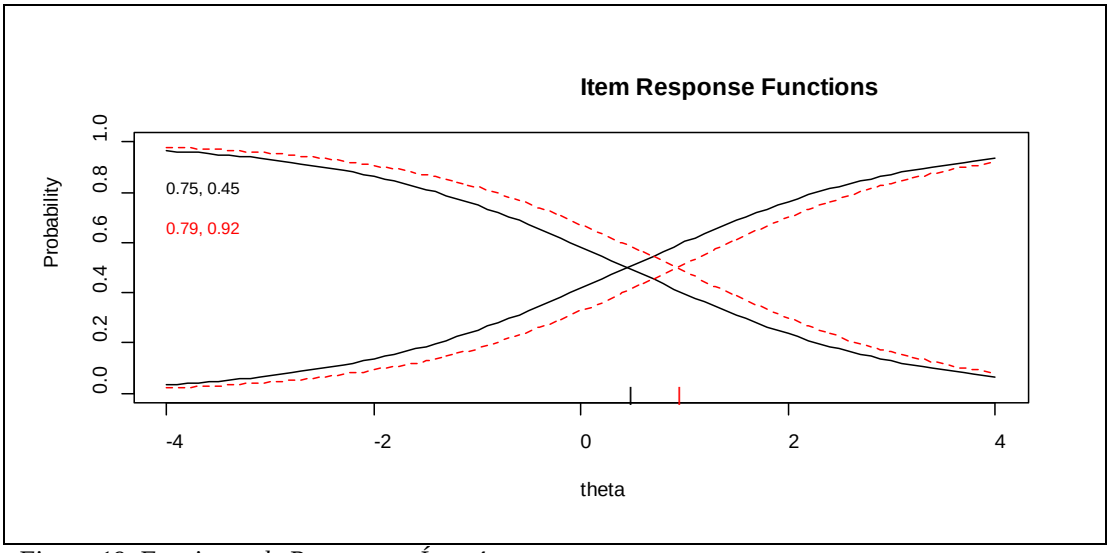


Figura 19. Funciones de Respuesta - Ítem 4

En la figura 20, podemos valorar las CCI atendiendo a los parámetros TRI estimados para cada grupo. Visualmente las CCI son muy semejantes cuando comparamos ambas versiones de la prueba.

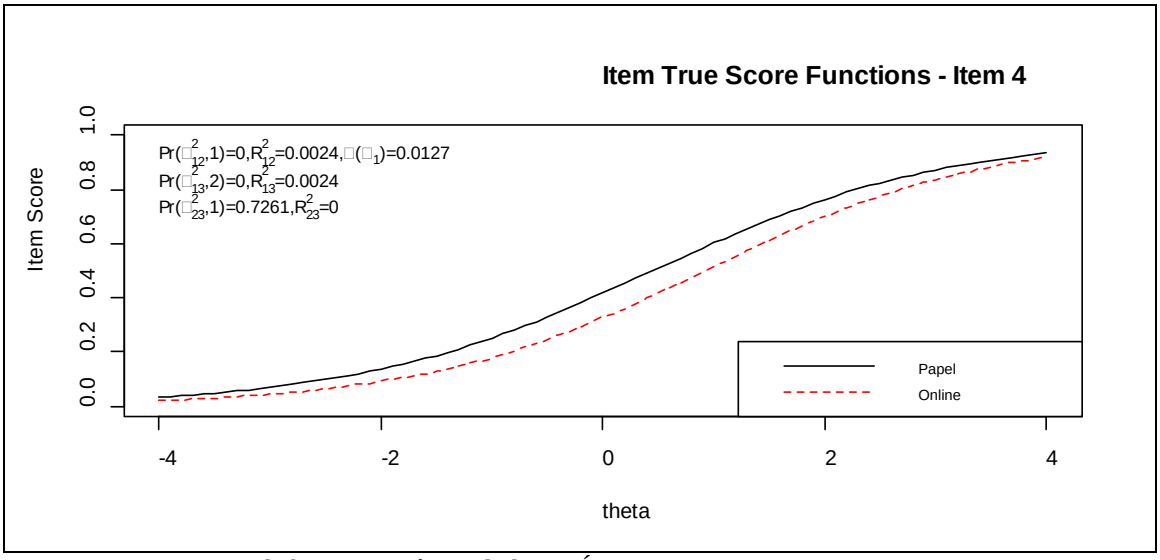


Figura 20. Funciones de la puntuación verdadera – Ítem 4

Junto a esta gráfica, podemos encontrar unos valores que hacen referencia a las pruebas $\chi 2$ para la diferencia entre los modelos; además se indican las diferencias en los valores Pseudo-R^2 y la diferencia entre los parámetros β_1. Si atendemos a estos valores, podemos observar cómo la prueba $\chi 2$ para la diferencia en los modelos 1 y 3,

muestra diferencias significativas (p<0,01), lo que nos indica la presencia de DVF en este ítem.

Los siguientes valores, nos van a permitir detectar de qué tipo de DVF estamos hablando. La prueba $\chi 2$ para la diferencia en los modelos 2 y 3 no muestra diferencias significativas (p>0,01), lo que nos indica que hay DVF uniforme, dado que favorece a los sujetos que realizan la prueba en papel independientemente del nivel de habilidad que tengan.

Por otro lado, la prueba $\chi 2$ para la diferencia en los modelos 1 y 2, muestra diferencias significativas (p<0,01), lo que significa que no hay DVF no uniforme.

En lo que se refiere a la medida del efecto debemos atender a los valores pseudo R^2 (R^2_{12}: 0,0024; R^2_{13}: 0,0024; R^2_{23}: 0). En todos los casos, los valores son inferiores a 0,035, por lo que podemos hablar de un ítem con DVF insignificante.

En la figura 21 se recoge la diferencia absoluta entre las CCI de los dos grupos. Esta diferencia, como ya hemos mencionado es mínima, aunque se hace más notoria entre los sujetos con niveles de habilidad intermedios.

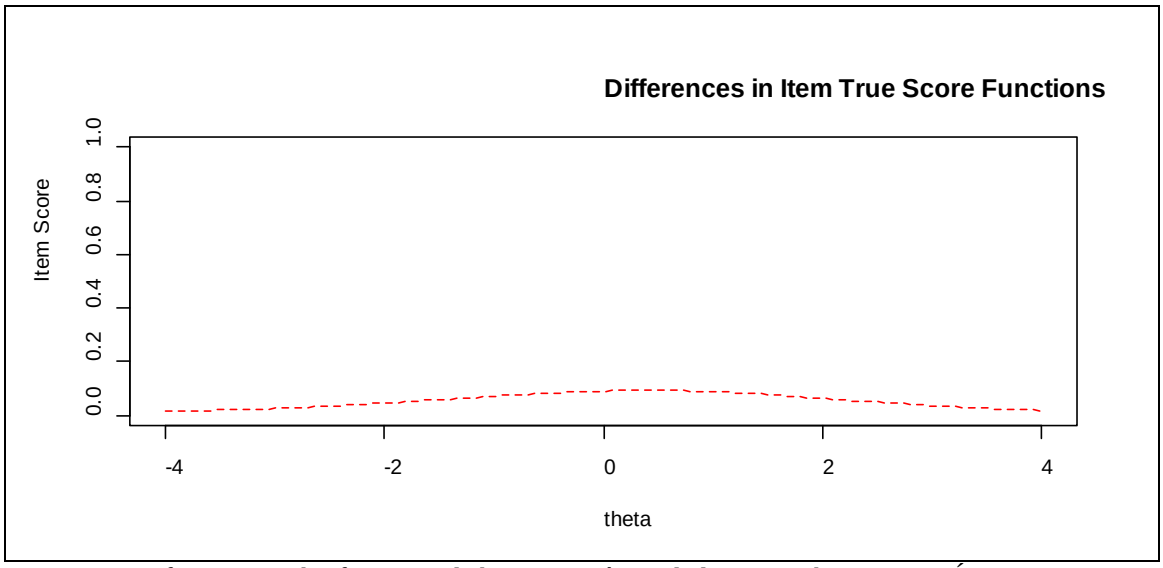


Figura 21. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 4

En la figura 22 se muestra el impacto del DVF; esto, unido a los valores arrojados en la gráfica anterior (R^2_{12}: 0,0024; R^2_{13}: 0,0024; R^2_{23}: 0;), nos permiten detectar cómo el impacto es mínimo, lo que nos da evidencias de que estamos ante un ítem con DVF irrelevante o insignificante.

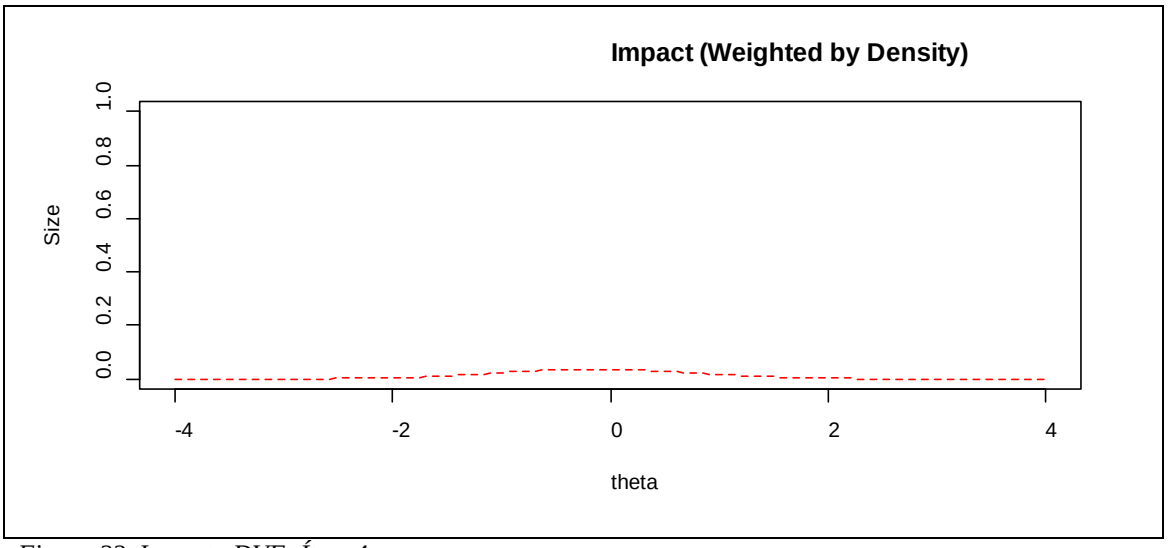


Figura 22. Impacto DVF- Ítem 4

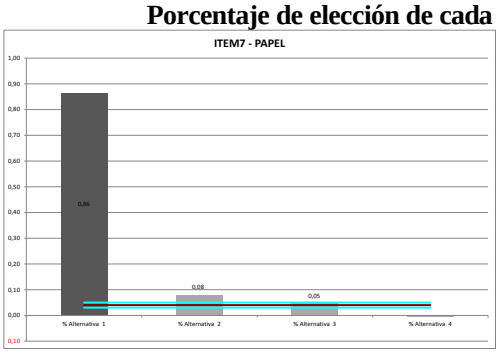
A continuación, realizamos un análisis en detalle de otro de los ítems que presenta DVF. En la tabla 7.30, se recogen las características fundamentales del ítem 7.

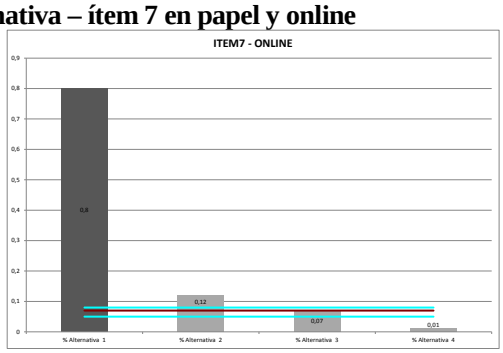
Tabla 7.30
Características y Funcionamiento Diferencial de Versiones en el Ítem 7

Descripción desde la TCT

Ítem 7	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,86	0,12	0,34	0,30	0,24	1	23,0	18,2	0,86	0,344
Online	0,80	0,16	0,40	0,40	0,33	1	21,6	16,0	0,80	0,402

Porcentaje de elección de cada alternativa – ítem 7 en papel y online





Modelo TRI de 2 Parámetros

Ítem 7	Parámetro a	Parámetro b	p
Papel	0,872	-2,317	0,014
Online	1,108	-1,527	0,112

Técnicas detección DVF Ítem 7

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,017	0,002	0,007	No Uniforme	0,001	0,0024	0,0013	Débil					

*Diferencias significativas ($p \leq 0,0021$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

Presentadas las características del ítem 7, podemos apreciar como los sujetos que realizan el ítem en papel, de nuevo obtienen una media superior a los que realizan la prueba online; estamos ante un ítem muy fácil, siendo ligeramente más fácil para los que lo realizan en papel. La correlación Biserial puntual corregida es superior en la prueba online (0,33) que en la prueba en papel (0,24) y de nuevo es un ítem que tiene una discriminación baja.

El modelo TRI de dos parámetros nos indica de nuevo que, tanto el parámetro de discriminación como el parámetro de dificultad, son superiores en la prueba online.

En la figura 23, se muestran las funciones de respuesta al ítem para ambos grupos, atendiendo a las estimaciones de los parámetros.

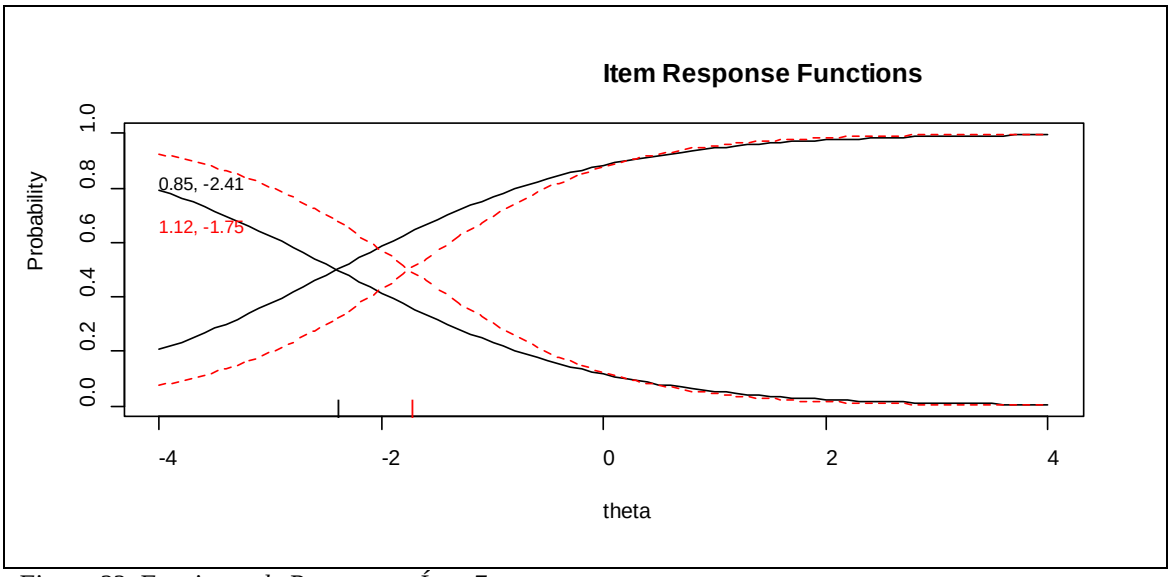


Figura 23. Funciones de Respuesta - Ítem 7

En lo que respecta a las técnicas de detección del DVF, podemos observar cómo además de la técnica de regresión logística, el método de Rajú, Lord y MH detectan el ítem 7 con DVF. Mientras que el método T.I.D. y Standard no detectan DVF en este ítem.

La figura 24, resume algunas características que nos permiten conocer en profundidad ante qué tipo de DVF nos encontramos. Los valores que aparecen en la gráfica nos permiten observar la significatividad ($p<0,01$) en la prueba $\chi 2$ para la diferencia en los modelos 1 y 3, lo que nos indica que estamos ante un ítem con DVF.

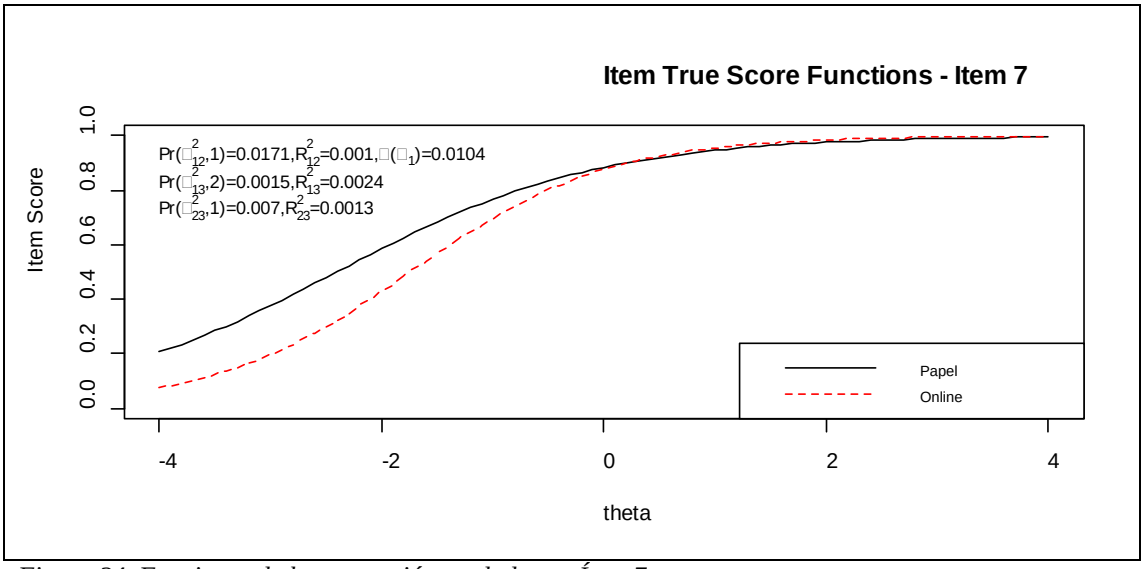


Figura 24. Funciones de la puntuación verdadera – Ítem 7

Concretamente, estamos ante un ítem con DVF no uniforme, ya que la prueba $\chi 2$ para la diferencia en los modelos 2 y 3 muestra diferencias significativas ($p<0,01$). Principalmente, estas diferencias se pronuncian más entre los sujetos con habilidades inferiores (ver figura 25).

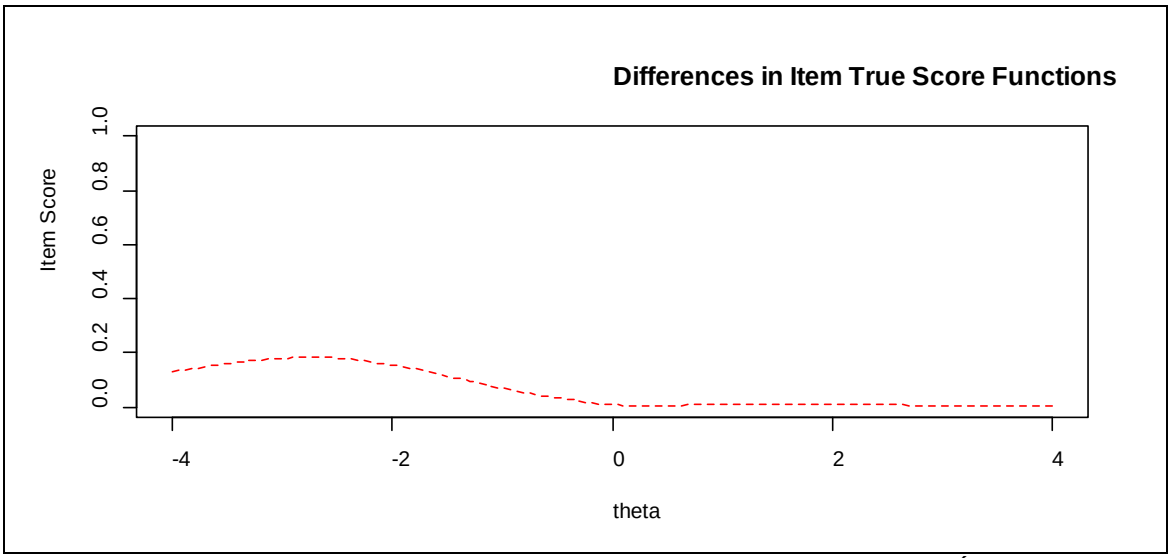


Figura 25. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 7

A pesar de que estas pruebas confirmen la existencia del ítem con DVF, el impacto es mínimo. La medida del efecto ($R^2 < 0,035$): (R^2_{12}: 0,001; R^2_{13}: 0,0024; R^2_{23}: 0,0013) nos permite hablar de un ítem con DVF insignificante (ver figura 26).

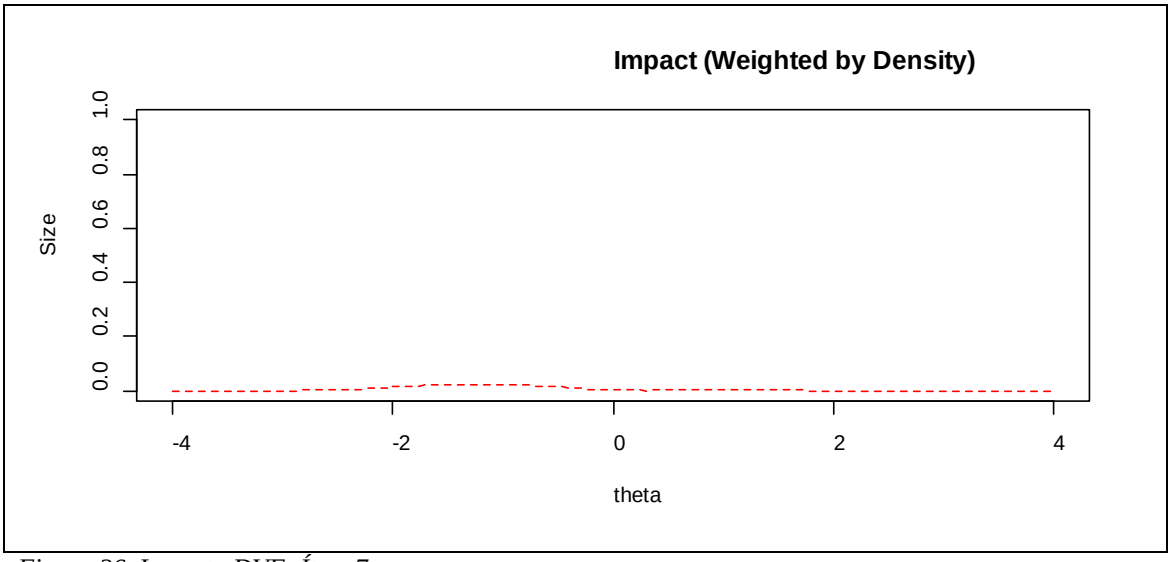


Figura 26. Impacto DVF- Ítem 7

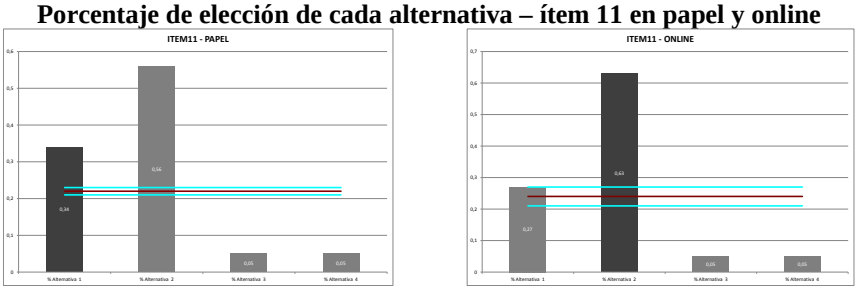
En las siguientes líneas, realizamos un análisis en detalle del ítem 11 que presenta DVF. En la tabla 7.31, se recogen las características fundamentales del ítem.

Tabla 7.31.
Características y Funcionamiento Diferencial de Versiones en el Ítem 11

Descripción desde la TCT

Ítem 11	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,34	0,23	0,47	0,24	0,16	1	24,1	21,4	0,36	0,479
Online	0,27	0,20	0,44	0,12	0,05	1	21,6	20,0	0,27	0,443

Porcentaje de elección de cada alternativa – ítem 11 en papel y online



Modelo TRI de 2 Parámetros

Ítem 11	Parámetro a	Parámetro b	p
Papel	0,365	1,611	0,100
Online	0,104	9,762	0,249

Técnicas detección DVF Ítem 11

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	DVF
0,000	0,000	0,000	No Uniforme	0,0021	0,0042	0,0021	Débil					

*Diferencias significativas ($p \leq 0{,}0032$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 11, se caracteriza por favorecer a los estudiantes que realizan la prueba en papel, dado que obtienen una media superior a la versión online. Es un ítem difícil y además lo es más para los sujetos que los que realizan la versión online. La correlación biserial puntual corregida, nos demuestra que este ítem tiene una discriminación muy baja.

En lo que respecta al modelo de TRI de dos parámetros, podemos observar, en la misma línea que los resultados obtenidos en la TCT, que el parámetro de dificultad es superior en la prueba online, mientras que el índice de discriminación es superior en

la prueba en papel. En la figura 27, se recogen las funciones de respuesta al ítem para ambos grupos, atendiendo a las estimaciones de los parámetros. En ella podemos observar algunas diferencias que debemos estudiar.

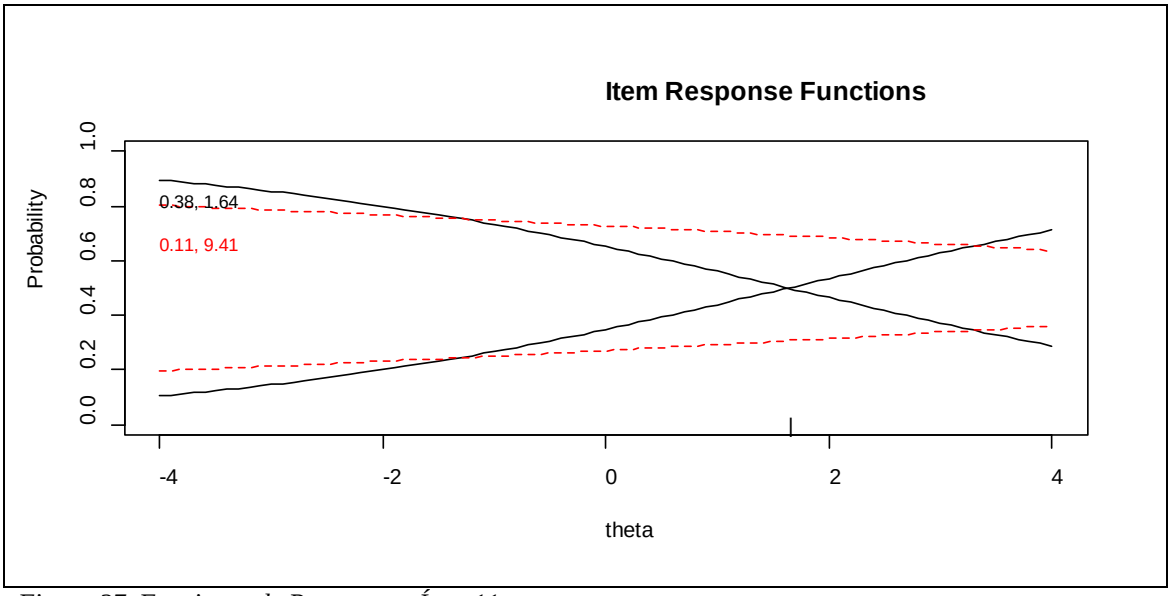


Figura 27. Funciones de Respuesta - Ítem 11

Como en los casos anteriores, además de detectar el DVF con el procedimiento de Regresión Logística, se han utilizado otros métodos que también han detectado que este ítem presenta DVF, es el caso del método MH. En cambio, el método T.I.D, Lord, Standard y Rajú, no consideran que este ítem presente DVF.

A continuación, en la figura 28 se lleva a cabo un estudio más profundo sobre el tipo de DVF detectado.

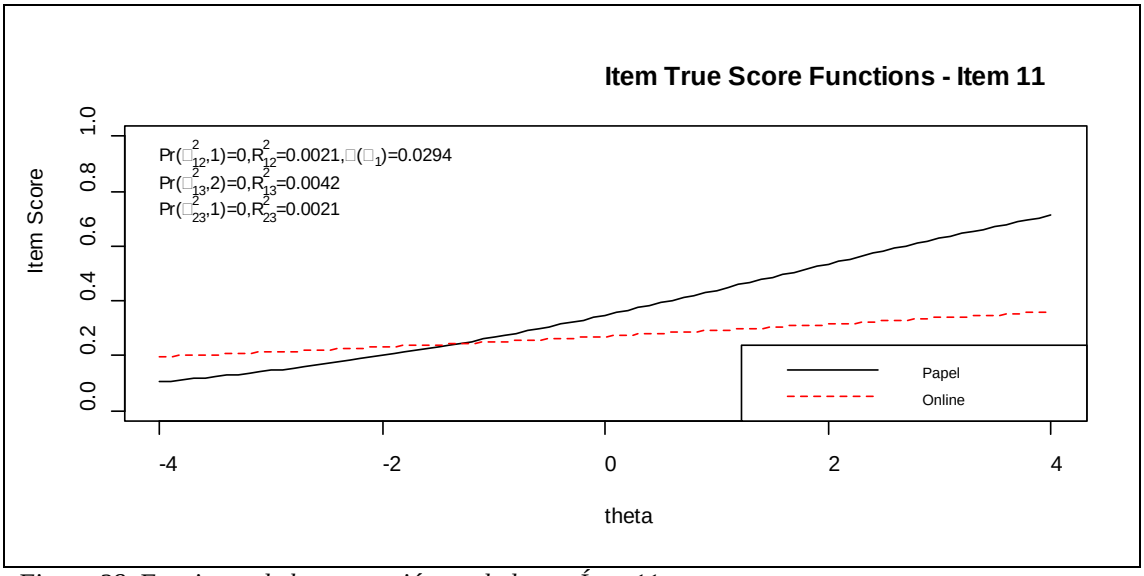


Figura 28. Funciones de la puntuación verdadera – Ítem 11

La significatividad (p<0,000) de la prueba $\chi 2$ para la diferencia en los modelos 1 y 3, nos indica que estamos ante un ítem con DVF. Concretamente se trata de un ítem con DVF no uniforme, puesto que la prueba $\chi 2$ para la diferencia en los modelos 2 y 3, muestra diferencias significativas (p<0,01).

En la figura 29, se puede observar de nuevo, como las diferencias se pronuncian más entre los sujetos con habilidades superiores.

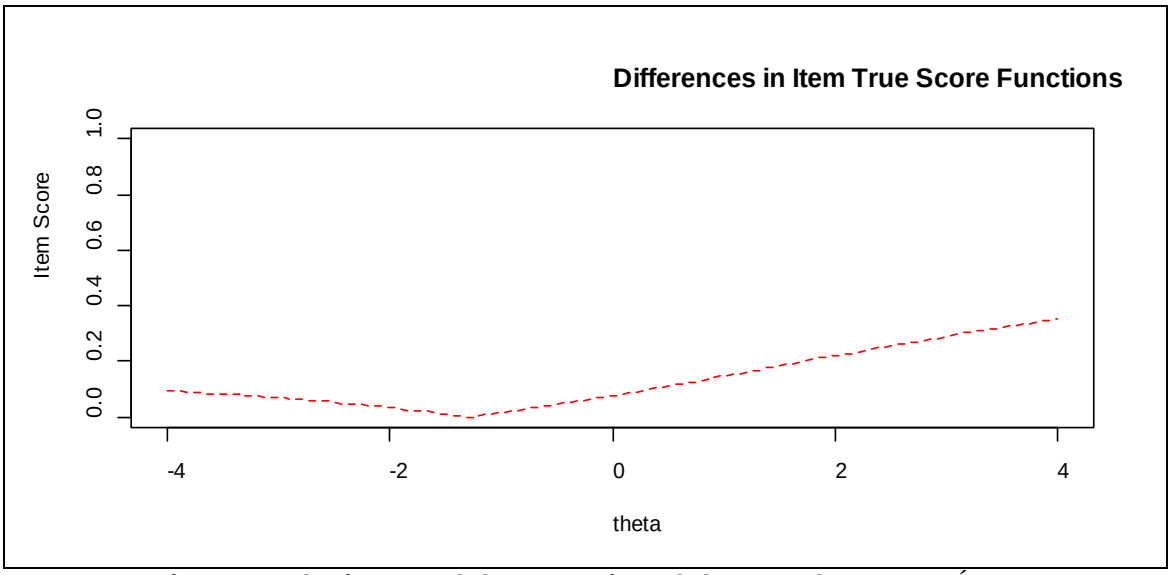


Figura 29. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 11

Ante estas evidencias empíricas, podemos considerar la existencia de DVF en el ítem 11, pero es necesario atender a la medida del efecto. Concretamente el impacto es mínimo puesto que ($R^2 < 0,035$): (R^2_{12}: 0,0021; R^2_{13}: 0,0042; R^2_{23}: 0,0021). Por ello, y al igual que en los casos anteriores, este ítem presenta DVF irrelevante (ver figura 30).

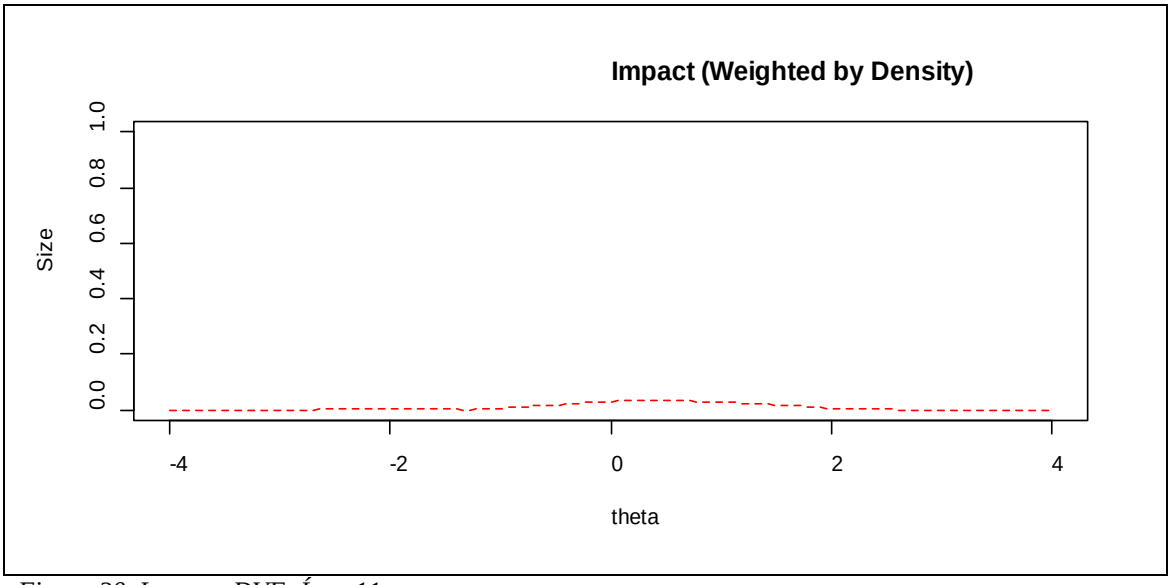


Figura 30. Impacto DVF- Ítem 11

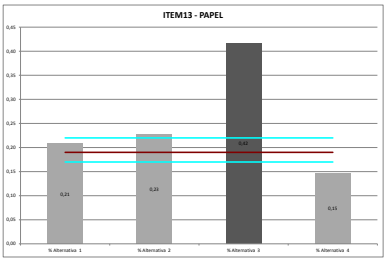
En lo que se refiere al siguiente ítem que presenta DVF, concretamente el ítem 13, en la tabla 7.32, se describe en detalle sus características.

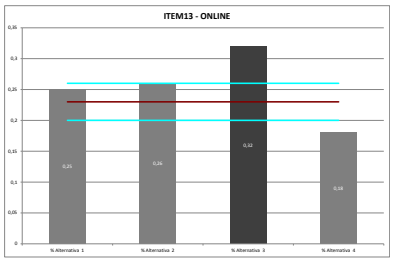
Tabla 7.32.
Características y Funcionamiento Diferencial de Versiones en el Ítem 13

Descripción desde la TCT

Ítem 13	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,42	0,24	0,49	0,40	0,32	3	24,8	20,5	0,42	0,494
Online	0,32	0,22	0,46	0,26	0,19	3	22,6	19,4	0,32	0,465

Porcentaje de elección de cada alternativa – ítem 13 en papel y online





Modelo TRI de 2 Parámetros

Ítem 13	Parámetro a	Parámetro b	p
Papel	0,774	0,572	0,001
Online	0,498	1,634	2,632

Técnicas detección DVF Ítem 13

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	No DIF	No DIF	DIF
0,001	0,000	0,000	No Uniforme	0,0011	0,0039	0,0027	Débil					

*Diferencias significativas ($p \leq 0{,}0038$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

Al igual que los casos anteriores, el ítem 13 tiene mayor media en los estudiantes que realizan la prueba en papel que online. Este ítem puede considerarse difícil, y lo es más para la prueba online, de ahí que la media sea inferior en este grupo. La correlación biserial puntual además nos indica la baja discriminación de este ítem en concreto en la prueba online.

El modelo de dos parámetros de TRI, nos indica de nuevo que en la prueba online el ítem es más difícil y el parámetro de discriminación es superior en la prueba

en papel. Este ítem es uno de los que no presentaba un ajuste adecuado en ambos grupos. En la figura 31 se recogen las funciones de respuesta al ítem para ambos grupos atendiendo a las estimaciones de los parámetros.

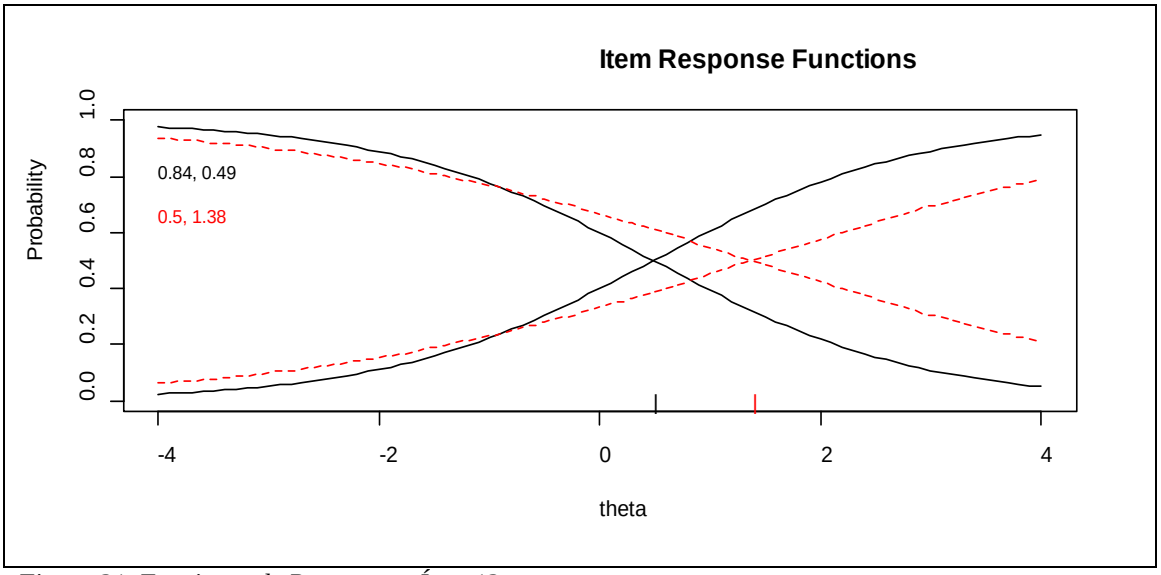


Figura 31. Funciones de Respuesta - Ítem 13

En lo que respecta al apartado de detección del DVF, además del método de Regresión Logística, los métodos Rajú, Lord y MH, también detectan este ítem con DVF. Mientras que los métodos T.I.D. y Standard no consideran que el ítem 13 presente DVF.

En la figura 32 podemos observar que el ítem 13 presenta DVF, dada la significatividad ($p<0,000$) de la prueba $\chi 2$ para la diferencia en los modelos 1 y 3. Como antes adelantábamos, es un ítem con DVF no uniforme (prueba $\chi 2$ para la diferencia en los modelos 2 y 3, muestra diferencias significativas, $p<0,01$).

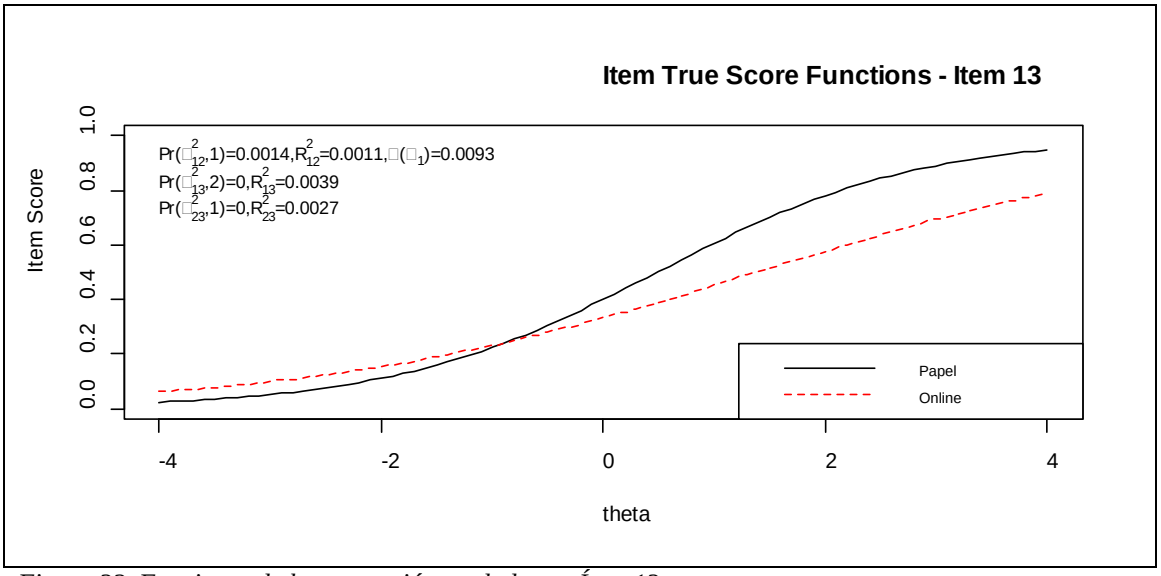


Figura 32. Funciones de la puntuación verdadera – Ítem 13

Reiteradamente, en la figura 33, se aprecia que las diferencias principalmente destacan entre los sujetos con habilidades más altas.

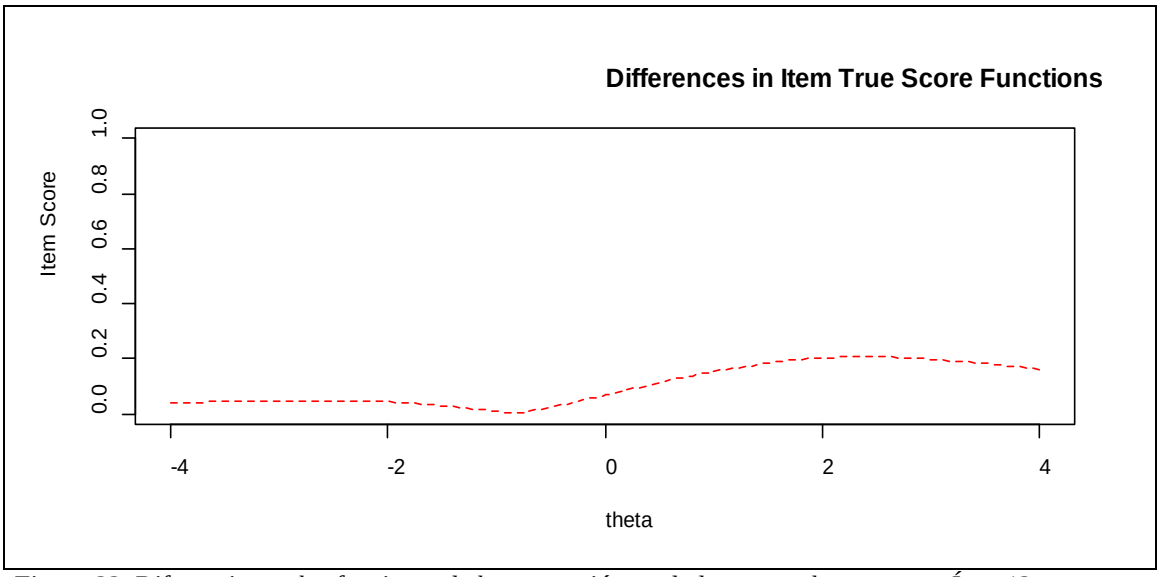


Figura 33. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 13

A pesar de estos resultados, la medida del efecto es muy pequeño, por lo que el impacto es mínimo: (R^2 < 0,035): (R^2_{12}: 0,0011; R^2_{13}: 0,0039; R^2_{23}: 0,0027), por ello y al igual que en los casos anteriores, este ítem presenta DVF irrelevante (ver figura 34).

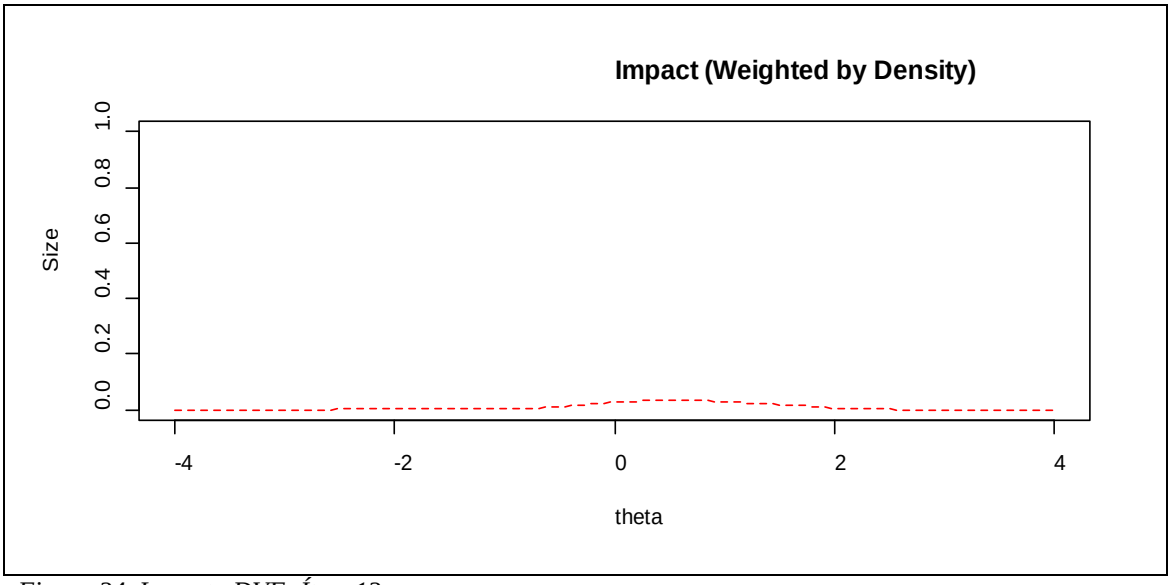


Figura 34. Impacto DVF- Ítem 13

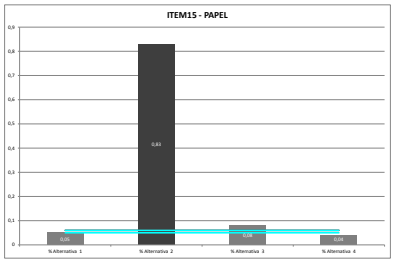
A continuación se muestran las características relativas al ítem 15 que presenta DVF. En la tabla 7.33 podemos observar las características del ítem.

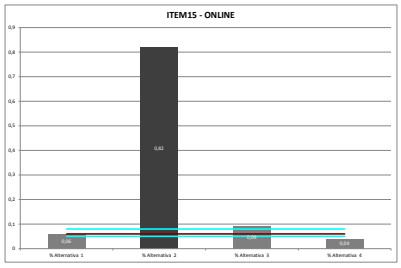
Tabla 7.33.
Características y Funcionamiento Diferencial de Versiones en el Ítem 15

Descripción desde la TCT

Ítem 15	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,83	0,14	0,38	0,42	0,36	2	23,4	17,4	0,83	0,378
Online	0,82	0,15	0,39	0,40	0,34	2	21,5	15,7	0,82	0,388

Porcentaje de elección de cada alternativa – ítem 15 en papel y online





Modelo TRI de 2 Parámetros

Ítem 15	Parámetro a	Parámetro b	p
Papel	1,236	-1,559	0,498
Online	1,220	-1,545	0,615

Técnicas detección DVF Ítem 15

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,001	0,002	0,840	Uniforme	0,002	0,002	0,000	Débil					

*Diferencias significativas ($p \leq 0{,}0044$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 15, como se puede observar en la tabla 7.33, tiene unas características semejantes en ambos grupos, con valores equivalentes en las medias y en la dificultad. Estamos ante un ítem muy fácil y que discrimina muy poco.

Además, los parámetros estimados según el modelo de 2 parámetros son de nuevo semejantes. En la figura 35 quedan recogidas las respuestas al ítem en ambos grupos.

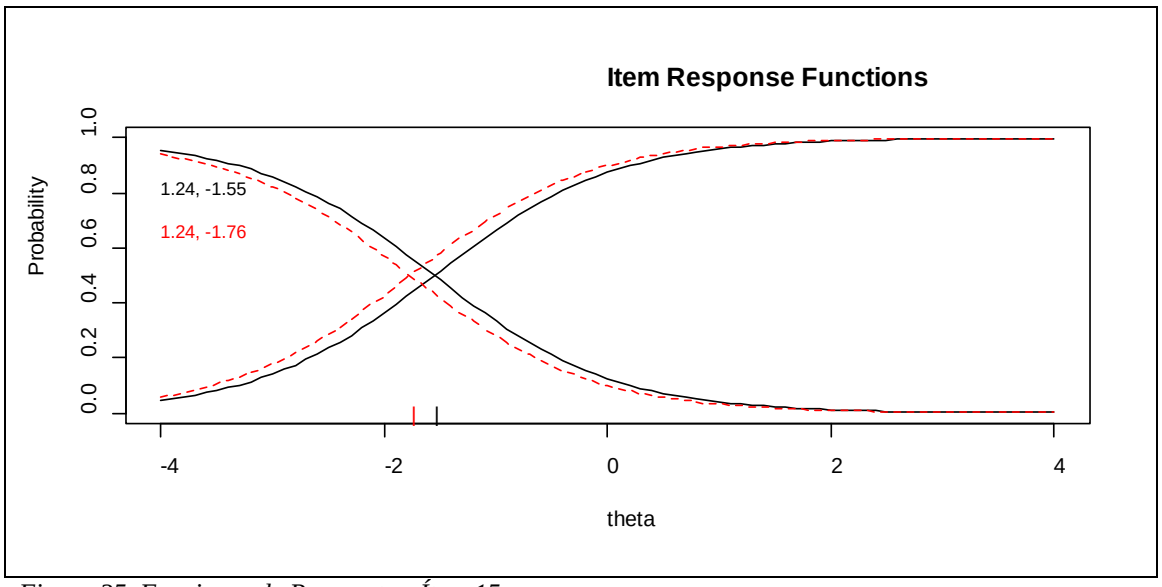


Figura 35. Funciones de Respuesta - Ítem 15

El estudio del DVF, demuestra que los métodos de Regresión Logística, así como el método de Rajú, el de Lord y MH detectan el ítem 15 con DVF mientras que los métodos T.I.D y Standard no.

Si llevamos a cabo un estudio más detallado, en la figura 36, observamos como la prueba $\chi 2$ para la diferencia en los modelos 1 y 3 es significativa y por tanto nos indica la presencia de DVF en el ítem 15. Además y más concretamente DVF no uniforme (la prueba $\chi 2$ para la diferencia en los modelos 2 y 3 no indica diferencias significativas, p>0,01).

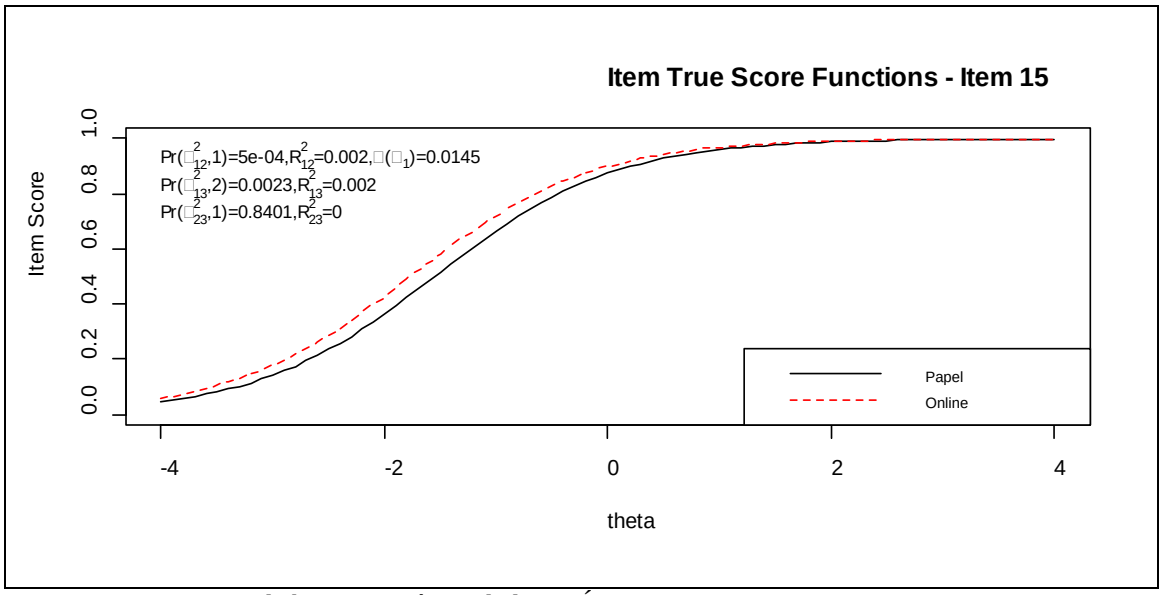


Figura 36. Funciones de la puntuación verdadera – Ítem 15

La figura 37, demuestra que estas diferencias son más acusadas en los sujetos con habilidades inferiores.

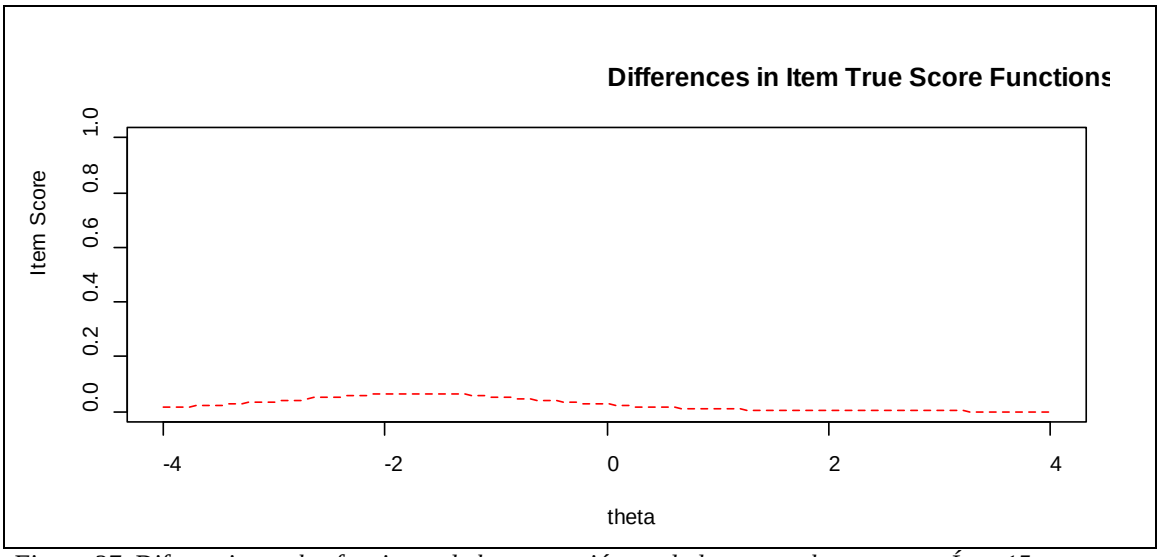


Figura 37. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 15

Ante este panorama no podemos obviar para tomar una decisión adecuada, la medida del efecto. Como podemos observar en la figura 38 y en los datos obtenidos (R^2 < 0,035) (R^2_{12}: 0,002; R^2_{13}: 0,002; R^2_{23}: 0,000), el impacto y la medida del efecto es mínimo, por ello y al igual que en casos anteriores, nos encontramos con un ítem que presenta DVF insignificante.

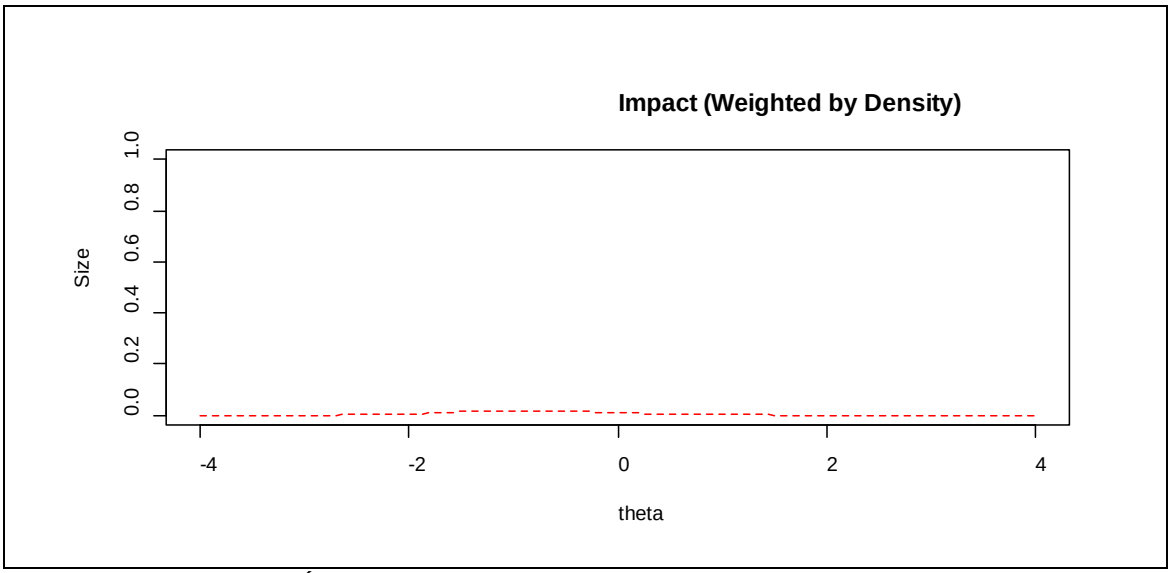


Figura 38. Impacto DVF- Ítem 15

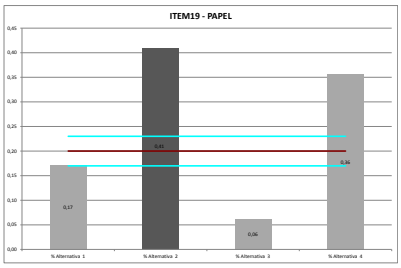
En lo que se refiere al siguiente ítem que presenta DVF, concretamente el ítem 19, en la tabla 7.34, se describe en detalle sus características.

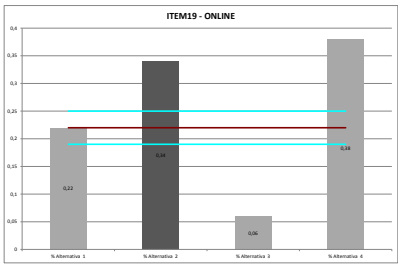
Tabla 7.34.
Características y Funcionamiento Diferencial de Versiones en el Ítem 19

Descripción desde la TCT

Ítem 19	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,41	0,24	0,49	0,28	0,19	2	24,1	21,1	0,42	0,493
Online	0,34	0,23	0,47	0,22	0,14	2	22,2	19,5	0,34	0,475

Porcentaje de elección de cada alternativa – Ítem 19 en papel y online





Modelo TRI de 2 Parámetros

Ítem 19	Parámetro a	Parámetro b	p
Papel	0,462	0,742	0,072
Online	0,337	1,978	0,889

Técnicas detección DVF Ítem 19

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	No DIF	No DIF	No DIF
0,015	0,003	0,017	No Uniforme	0,0007	0,0013	0,0006	Débil					

*Diferencias significativas (p≤0,0056) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 19, como apreciamos en la tabla 7.34 presenta una media superior en el grupo que realiza la prueba en papel, siendo el ítem más fácil para este grupo. Este ítem puede clasificarse como difícil además de tener poco poder de discriminación.

Continuando con los resultados ofrecidos por el modelo de TRI de 2 parámetros, podemos observar en la figura 39 el comportamiento de este ítem.

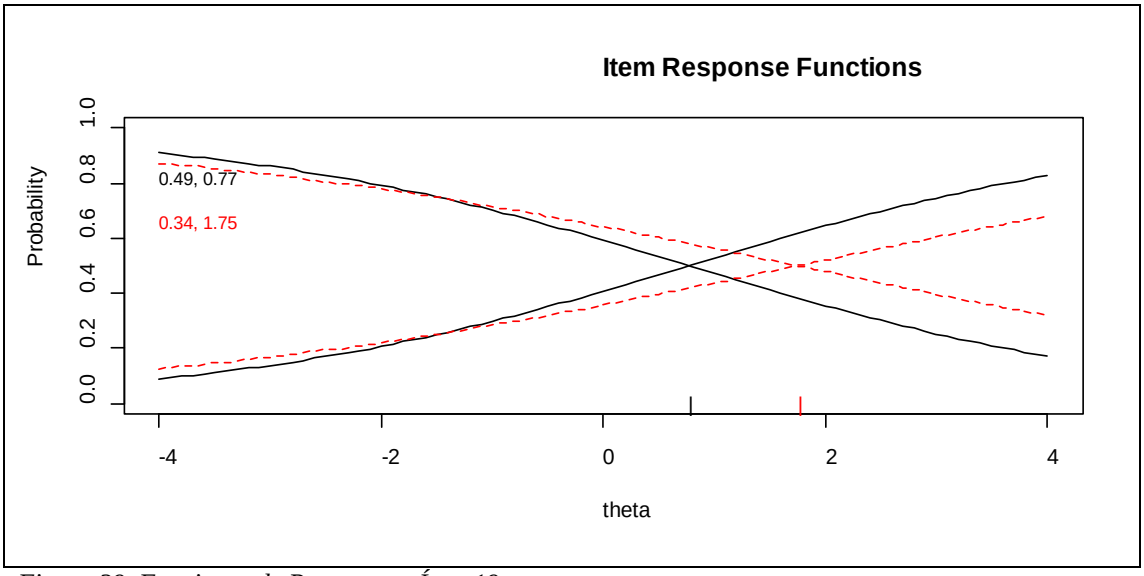


Figura 39. Funciones de Respuesta - Ítem 19

El caso de este ítem es diferente, puesto que los procedimientos para la detección del DVF (T.I.D, Standard, Rajú, Lord y MH) en ningún caso indican que el ítem 19 presente DVF.

En la figura 40, se puede apreciar la presencia de DVF, puesto que la prueba χ2 para la diferencia en los modelos 1 y 3 es significativa. Pero en lo que se refiere al tipo de DVF y atendiendo a los datos, la prueba χ2 para la diferencia en los modelos 1 y 2 (0,015) y los modelos 2 y 3 (0,017) no indica diferencias significativas p>0,01. El programa detecta el ítem 19 con DVF, y como apreciamos en la figura se trata de DVF no uniforme.

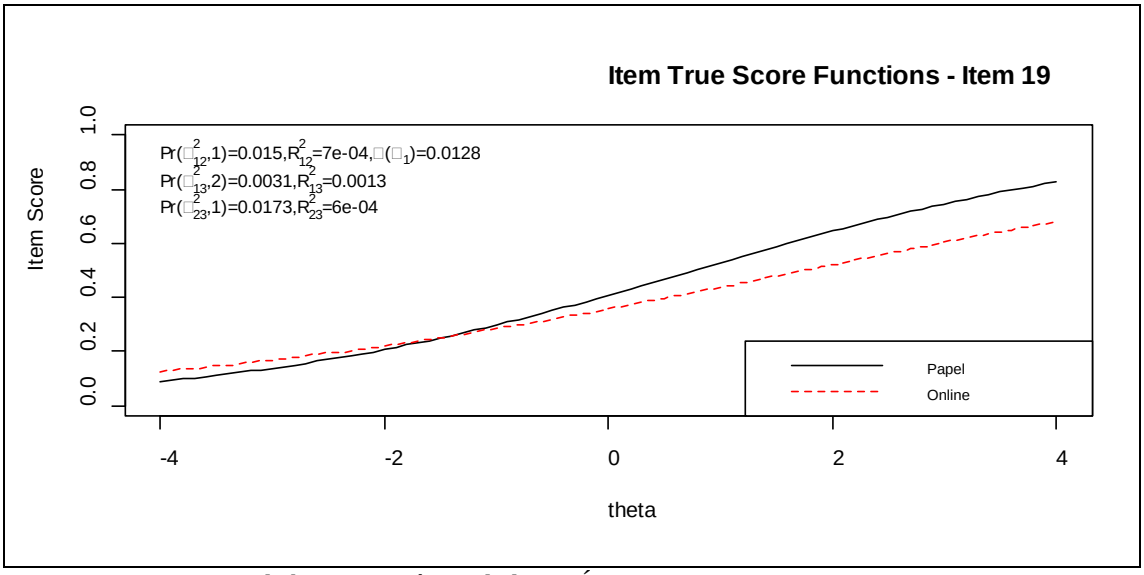


Figura 40. Funciones de la puntuación verdadera – Ítem 19

Las pequeñas diferencias que se observan nos indican que sucede más en los sujetos que tienen habilidades más altas (ver figura 41).

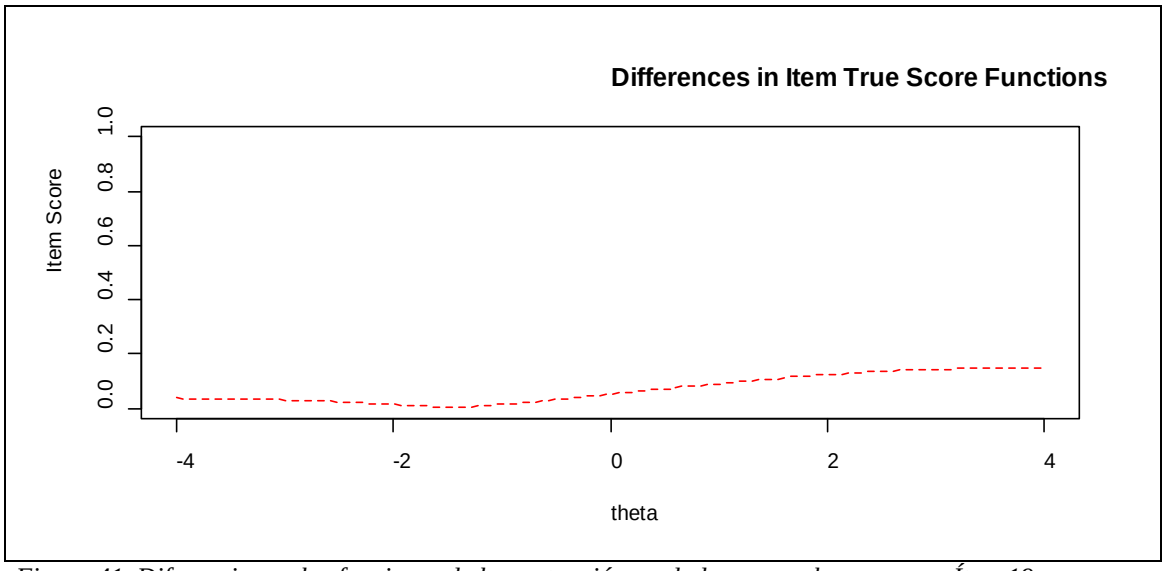


Figura 41. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 19

A pesar de detectar DVF, la medida del efecto nos indica un impacto muy débil: (R^2 < 0,035): (R^2_{12}: 0,0007; R^2_{13}: 0,0013; R^2_{23}: 0,0006). Por ello y como se refleja en la figura 42, el ítem 19 presenta de nuevo DVF irrelevante.

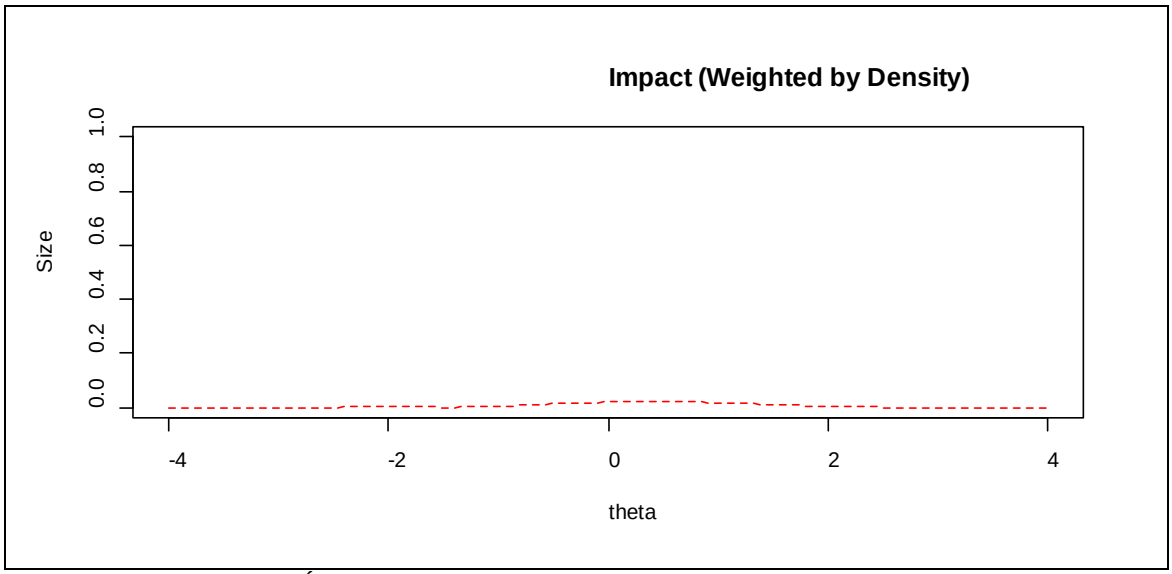


Figura 42. Impacto DVF- Ítem 19

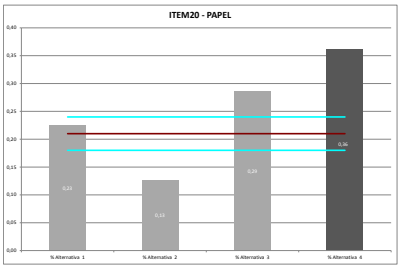
A continuación, realizamos un análisis en detalle de otro de los ítems que presenta DVF. En la tabla 7.35, se recogen las características fundamentales del ítem 20.

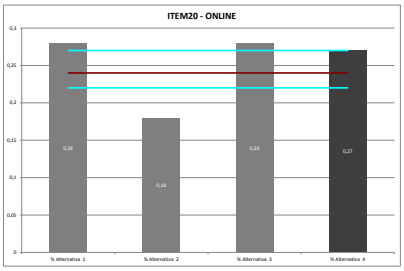
Tabla 7.35.
Características y Funcionamiento Diferencial de Versiones en el Ítem 20

Descripción desde la TCT

Ítem 20	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,36	0,23	0,48	0,36	0,27	4	24,8	20,9	0,38	0,484
Online	0,27	0,19	0,44	0,31	0,24	4	23,4	19,4	0,27	0,442

Porcentaje de elección de cada alternativa – ítem 20 en papel y online





Modelo TRI de 2 Parámetros

Ítem 20	Parámetro a	Parámetro b	p
Papel	0,842	0,730	1,497
Online	0,674	1,656	0,000

Técnicas detección DVF Ítem 20

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,000	0,000	0,016	Uniforme	0,0018	0,0024	0,0007	Débil					

*Diferencias significativas ($p \leq 0{,}0059$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

En lo que concierne al ítem 20, podemos observar diferencias entre ambas versiones. El resultado evidencia que puntuaciones superiores en la prueba en papel, obteniendo mejor media y siendo para este grupo el ítem más fácil. Este ítem puede clasificarse como difícil además de tener poco poder de discriminación.

Atendiendo al modelo de 2 parámetros, este ítem era uno de los que no ajustaba adecuadamente en la prueba online, como puede observarse en la tabla 7.35 y en la figura 43 donde se puede valora su comportamiento.

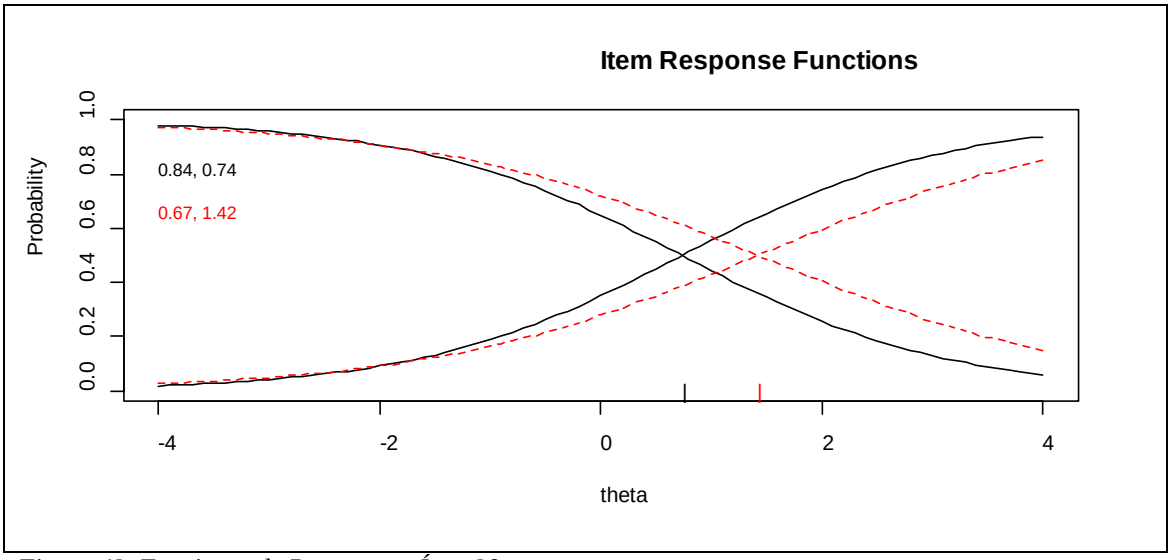


Figura 43. Funciones de Respuesta - Ítem 20

Los procedimientos para la detección del DVF que indican presencia del mismo en el ítem 20, son los métodos de Regresión Logística, Rajú, Lord y MH, mientras que los métodos T.I.D. y Standard indican que este ítem no presenta DVF.

En la figura 44, se puede apreciar la presencia de DVF, puesto que la prueba $\chi 2$ para la diferencia en los modelos 1 y 3 es significativa. Específicamente, DVF uniforme (la prueba $\chi 2$ para la diferencia en los modelos 2 y 3 no indica diferencias significativas, p>0,01).

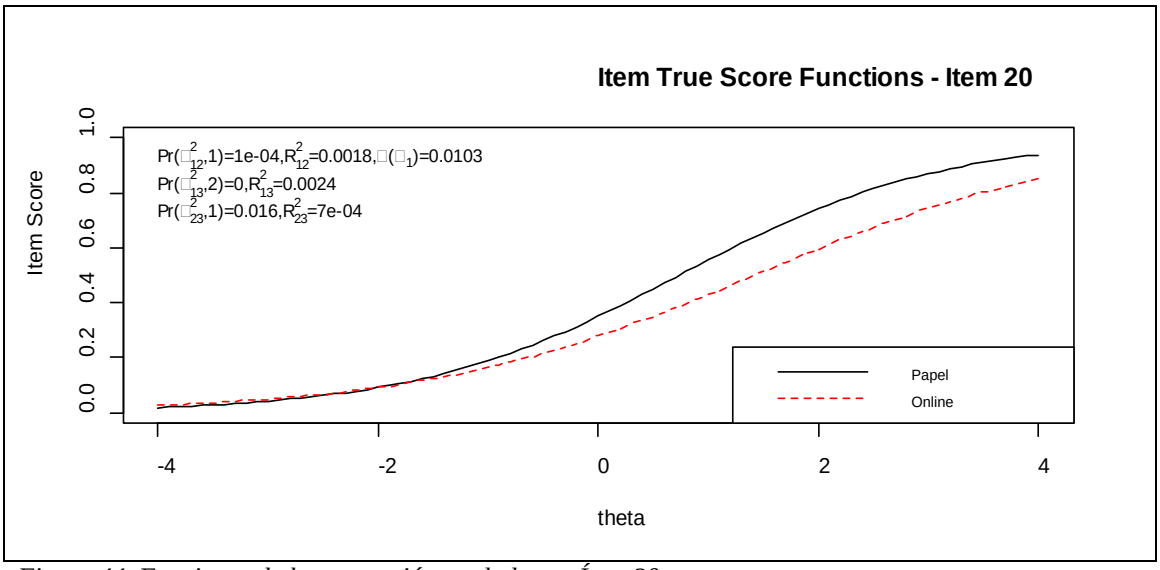


Figura 44. Funciones de la puntuación verdadera – Ítem 20

Las pequeñas diferencias que se observan nos indican que sucede en los sujetos que tienen habilidades más altas (ver figura 45).

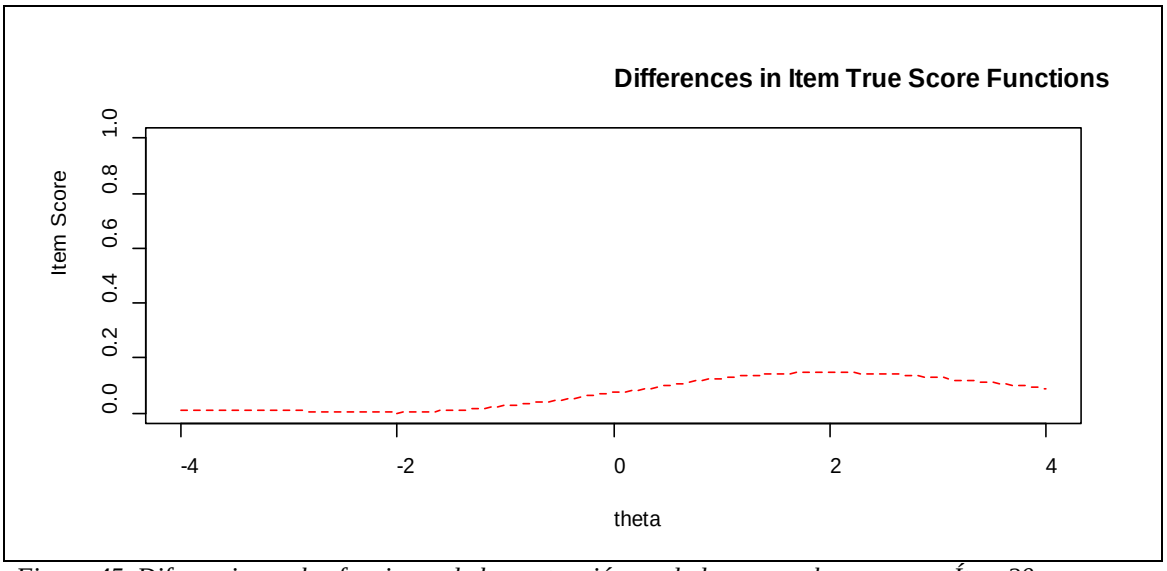


Figura 45. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 20

A pesar de detectar DVF, la medida del efecto nos indica un impacto muy débil: (R^2 < 0,035): (R^2_{12}: 0,0018; R^2_{13}: 0,0024; R^2_{23}: 0,0007). Por ello y como se refleja en la figura 46, el ítem 20 presenta DVF irrelevante.

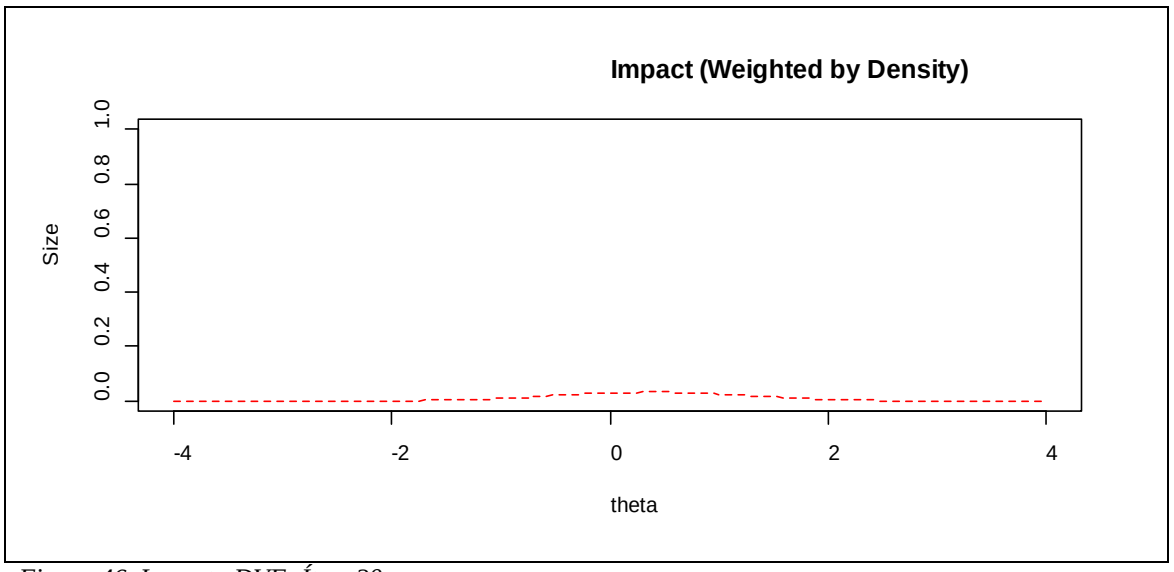


Figura 46. Impacto DVF- Ítem 20

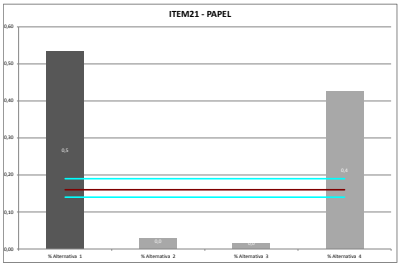
En lo que se refiere al siguiente ítem que presenta DVF, concretamente el ítem 21, en la tabla 7.36, se describe en detalle sus características.

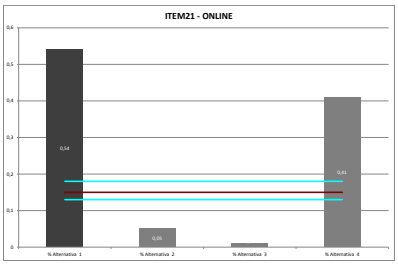
Tabla 7.36.
Características y Funcionamiento Diferencial de Versiones en el Ítem 21

Descripción desde la TCT

Ítem 21	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,53	0,25	0,50	0,27	0,18	1	23,7	20,8	0,52	0,500
Online	0,54	0,25	0,50	0,28	0,20	1	21,9	18,7	0,54	0,499

Porcentaje de elección de cada alternativa – ítem 21 en papel y online





Modelo TRI de 2 Parámetros

Ítem 21	Parámetro a	Parámetro b	p
Papel	0,375	-0,353	0,322
Online	0,463	-0,346	0,361

Técnicas detección DVF Ítem 21

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,002	0,007	0,983	Uniforme	0,0011	0,0011	0,0000	Débil					

*Diferencias significativas ($p \leq 0,0062$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 21, muestra que en ambas versiones tiene un comportamiento semejante, las medias son equivalentes; así como las correlaciones biseriales puntuales que demuestran una discriminación baja. Es de destacar este ítem como uno de los pocos que favorece a los sujetos que realizan la prueba online, su media el ligeramente superior (0,54) que a la alcanzada en la prueba en papel (0,52).

Lo mismo sucede cuando atendemos a la TRI, según el modelo de 2 parámetros, en ambas versiones los resultados son semejantes, pero el índice de discriminación y de dificultad es superior en la prueba online. En la figura 47 podemos observar el comportamiento de este ítem.

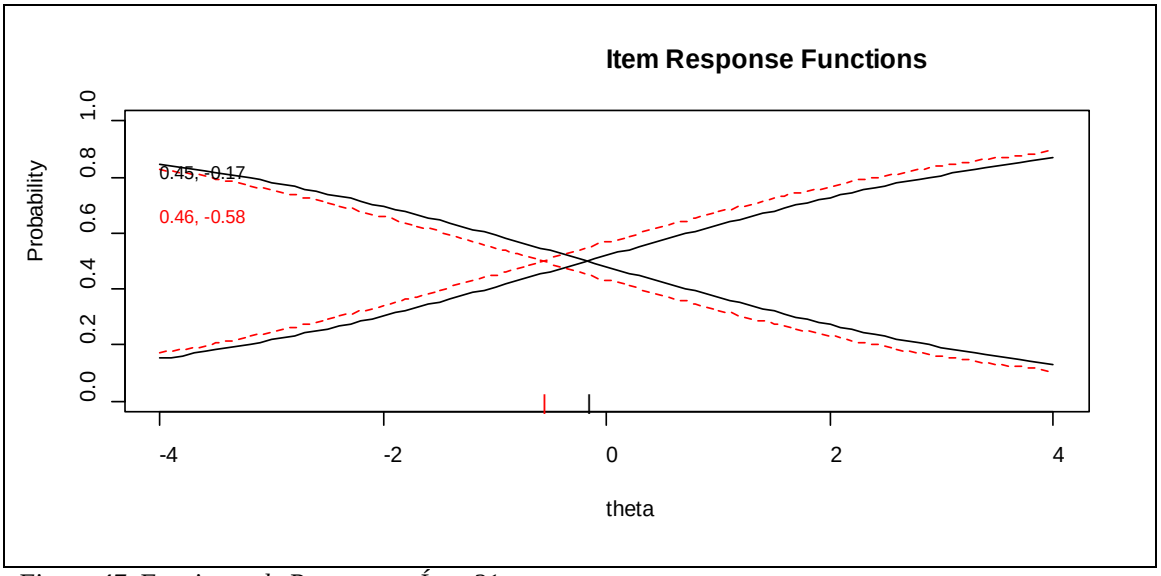


Figura 47. Funciones de Respuesta - Ítem 21

El ítem 21 es detectado como un ítem con DVF por los métodos de Regresión Logística, Rajú, Lord y MH. Mientras que los métodos de T.I.D. y Standard no detectan DVF en el ítem.

Según el procedimiento de Regresión Logística este ítem presenta DVF, porque la prueba $\chi 2$ para la diferencia en los modelos 1 y 3 es significativa. En la figura 48 podemos apreciar que se trata concretamente de un ítem con DVF uniforme, puesto que la prueba $\chi 2$ para la diferencia en los modelos 2 y 3 no indica diferencias significativas, p>0,01.

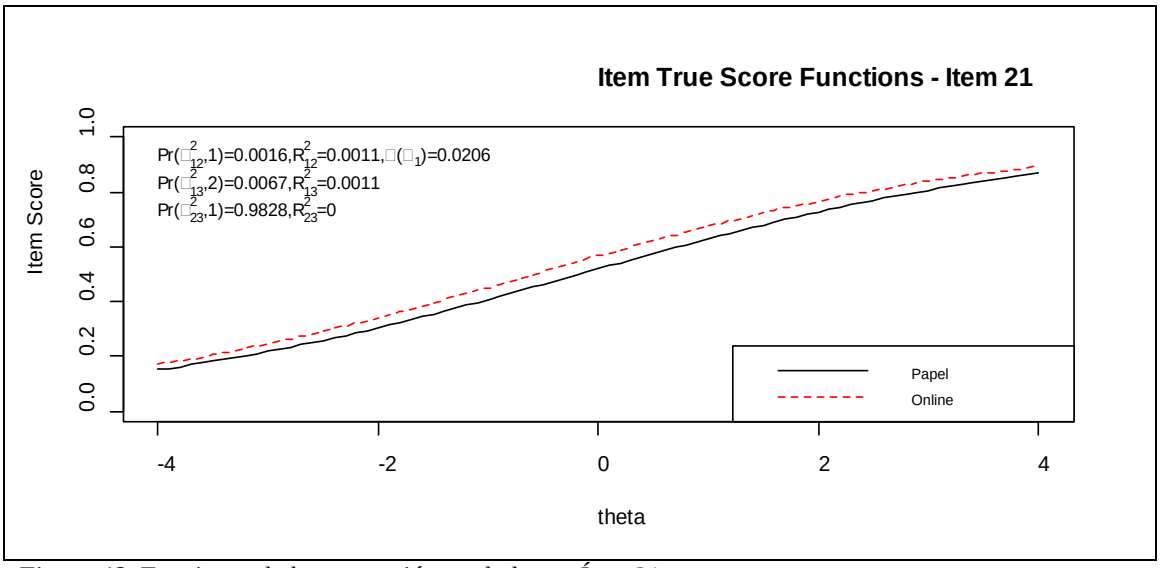


Figura 48. Funciones de la puntuación verdadera – Ítem 21

En la siguiente figura (49), podemos observar gráficamente las diferencias existentes entre ambos grupos, normalmente estas diferencias se observan principalmente en estudiantes con habilidades intermedias.

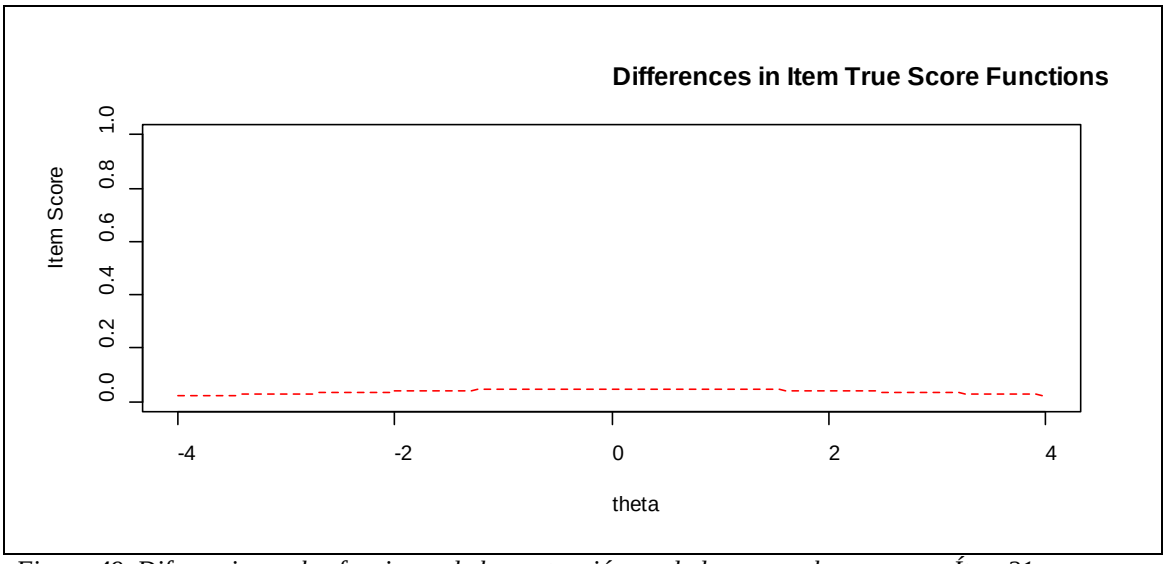


Figura 49. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 21

La medida del efecto es muy pequeño: (R^2 < 0,035): (R^2_{12}: 0,0011; R^2_{13}: 0,0011; R^2_{23}: 0,0000), así como la figura 50 nos muestra que el ítem 21 presenta DVF insignificante.

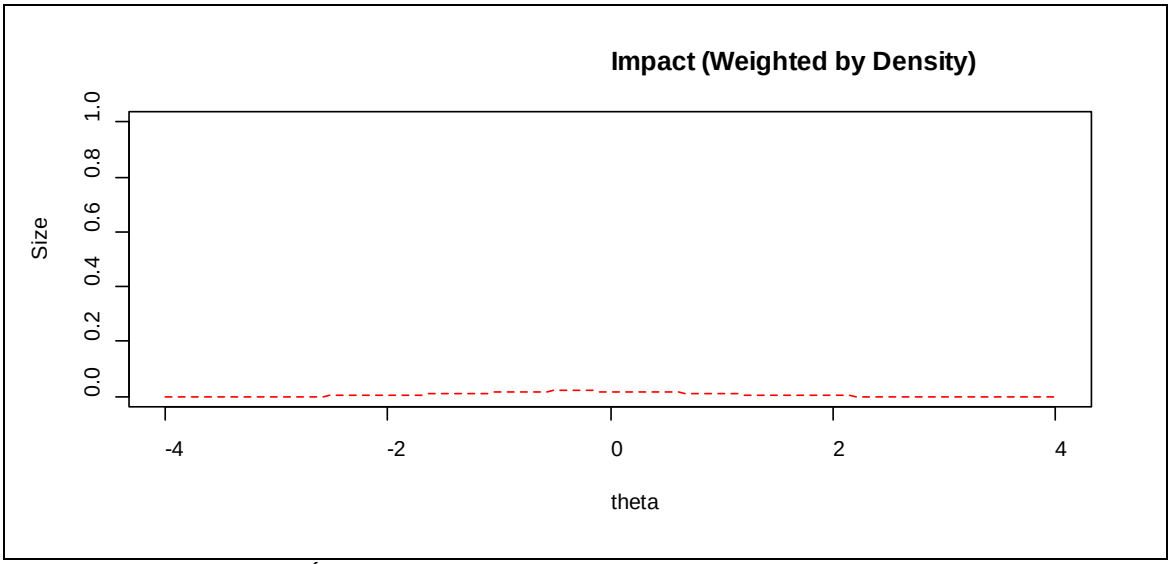


Figura 50. Impacto DVF- Ítem 21

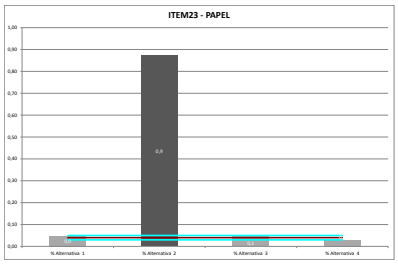
A continuación se muestran las características relativas al ítem 23 que presenta DVF. En la tabla 7.37 podemos observar las características del ítem.

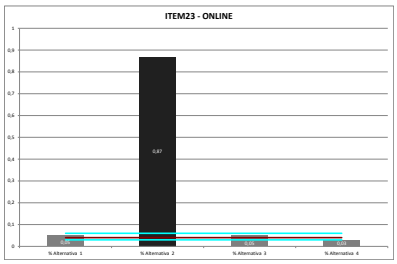
Tabla 7.37.
Características y Funcionamiento Diferencial de Versiones en el Ítem 23

Descripción desde la TCT

Ítem 23	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,87	0,11	0,33	0,38	0,33	2	23,1	17,0	0,87	0,341
Online	0,87	0,12	0,34	0,49	0,44	2	21,5	13,3	0,87	0,340

Porcentaje de elección de cada alternativa – ítem 23 en papel y online





Modelo TRI de 2 Parámetros

Ítem 23	Parámetro a	Parámetro b	p
Papel	1,324	-1,773	0,149
Online	2,350	-1,383	0,831

Técnicas detección DVF Ítem 23

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,000	0,000	0,000	No Uniforme	0,0052	0,011	0,0058	Débil					

*Diferencias significativas ($p \leq 0,0068$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 23 presenta unas características psicométricas equivalentes en ambos grupos. La media y la dificultad arrojan los mismo valores; la correlación biserial puntual parecida pero superior en la prueba online. Estamos ante un ítem muy fácil con una discriminación baja.

En lo que respecta al modelo de TRI de 2 parámetros, podemos observar que el índice de discriminación y de dificultad es superior en la prueba online. En la figura 51

se recogen las funciones de respuesta al ítem para ambos grupos atendiendo a las estimaciones de los parámetros.

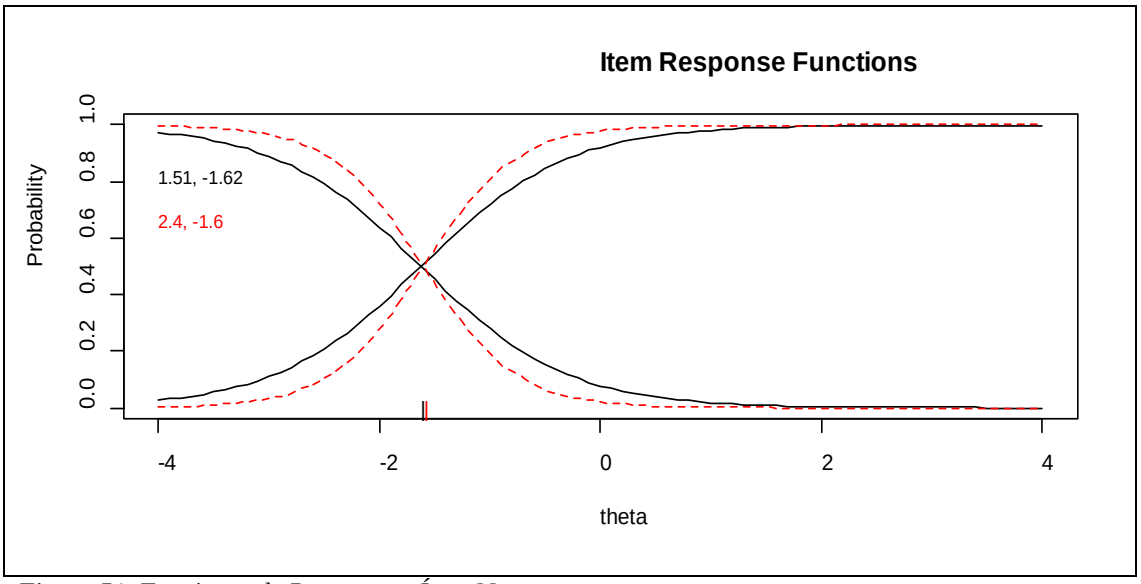


Figura 51. Funciones de Respuesta - Ítem 23

Los métodos de Regresión Logística, Rajú, Lord y MH, detectan el ítem 23 con presencia de DVF. Mientras que los métodos T.I.D. y Standard no consideran que el ítem 23 presente DVF.

El ítem 23 presenta DVF dado que en la prueba $\chi 2$ para la diferencia en los modelos 1 y 3, la significatividad es $p<0,000$. En la figura 52 podemos observarlo además de visualizar que se trata de un ítem con DVF no uniforme (prueba $\chi 2$ para la diferencia en los modelos 2 y 3, muestra diferencias significativas, $p<0,01$).

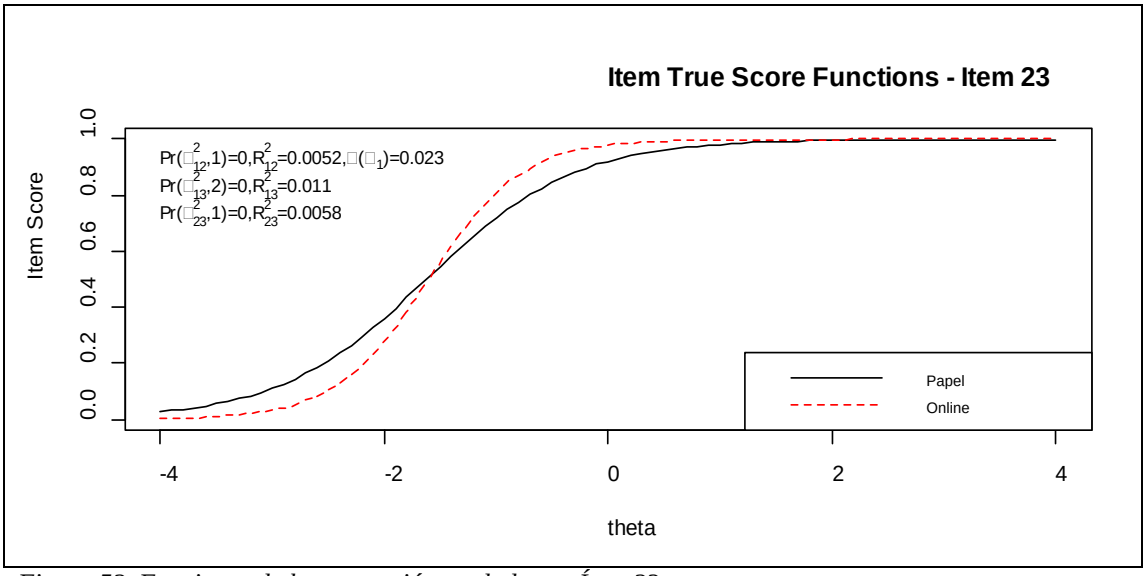


Figura 52. Funciones de la puntuación verdadera – Ítem 23

Las diferencias observables son pronunciadas en los casos en los que los sujetos tienen habilidades inferiores (ver figura 53).

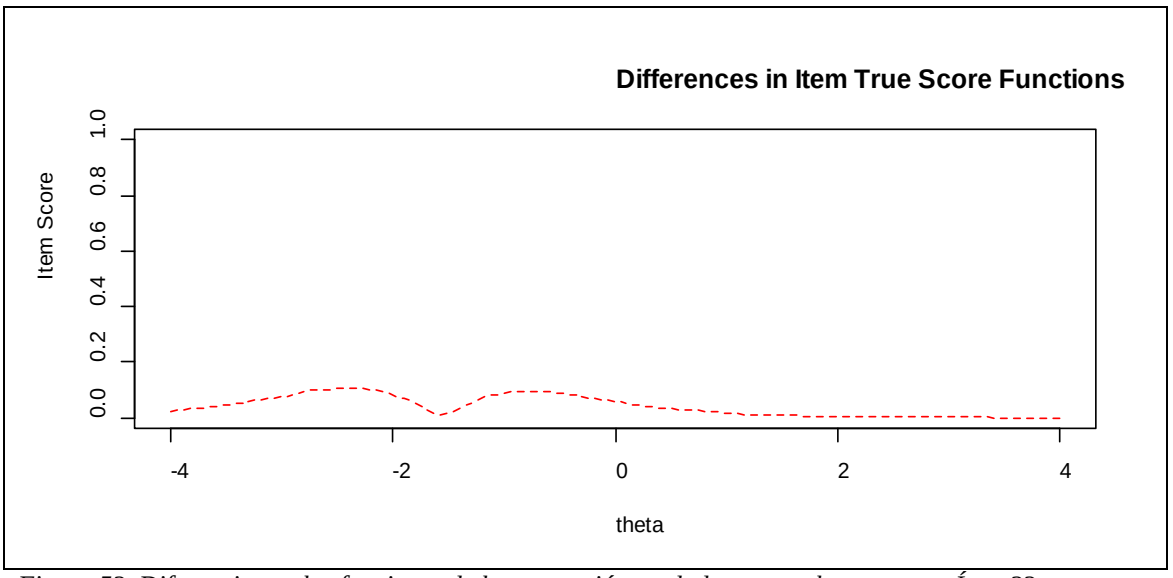


Figura 53. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 23

La medida del efecto alcanzado (R^2 < 0,035): (R^2_{12}: 0,0052; R^2_{13}: 0,011; R^2_{23}: 0,0058) podemos observar que es muy pequeña (ver figura 54). Por ello, el ítem 23 presenta DVF débil e irrelevante.

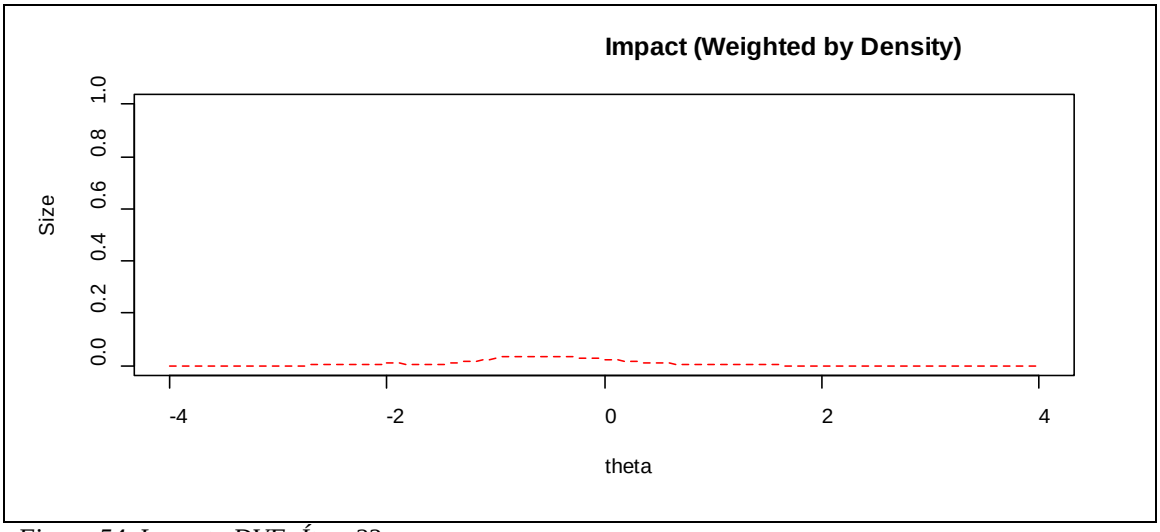


Figura 54. Impacto DVF- Ítem 23

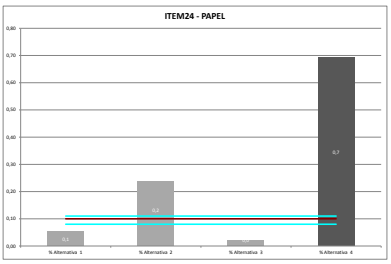
A continuación, presentamos en detalle (tabla 7.38) las características del ítem 24 y el estudio llevado a cabo del DVF.

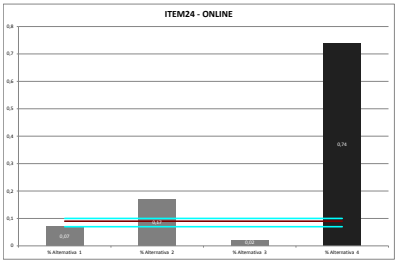
Tabla 7.38.
Características y Funcionamiento Diferencial de Versiones en el Ítem 24

Descripción desde la TCT

Ítem 24	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,69	0,21	0,46	0,25	0,17	4	23,2	20,3	0,70	0,460
Online	0,74	0,19	0,44	0,29	0,22	4	21,4	17,6	0,74	0,439

Porcentaje de elección de cada alternativa – ítem 24 en papel y online





Modelo TRI de 2 Parámetros

Ítem 24	Parámetro a	Parámetro b	p
Papel	0,657	-1,391	0,376
Online	0,575	-1,949	0,010

Técnicas detección DVF Ítem 24

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,000	0,000	0,430	Uniforme	0,0041	0,0042	0,0001	Débil					

*Diferencias significativas ($p \leq 0,0071$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 24 es uno de los pocos ítems que favorece a los sujetos que realizan la prueba online, la media y la facilidad es superior frente a los sujetos que realizan la prueba en papel. Es un ítem muy fácil y con baja discriminación.

En lo que se refiere a los parámetros de la TRI, en la figura 55 podemos observar el comportamiento de este ítem que tiende a ser semejante en ambos grupos.

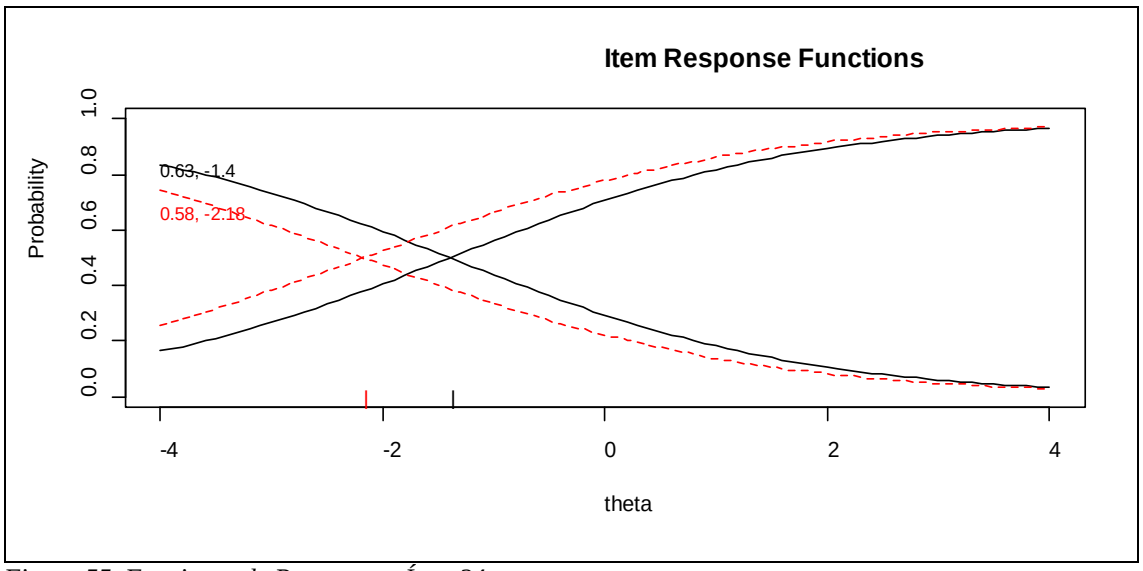


Figura 55. Funciones de Respuesta - Ítem 24

Al igual que sucedía anteriormente, los procedimientos que detectan el ítem 24 con DVF, son los métodos de Regresión Logística, Rajú, Lord y MH. Mientras que los métodos de T.I.D. y Standard no detectan DVF en el ítem.

En la figura 56, podemos observar que según el procedimiento de Regresión Logística, el ítem 24 presenta DVF (la prueba $\chi 2$ para la diferencia en los modelos 1 y 3 es significativa). Además estamos ante un ítem Uniforme (prueba $\chi 2$ para la diferencia en los modelos 2 y 3 no indica diferencias significativas, p>0,01).

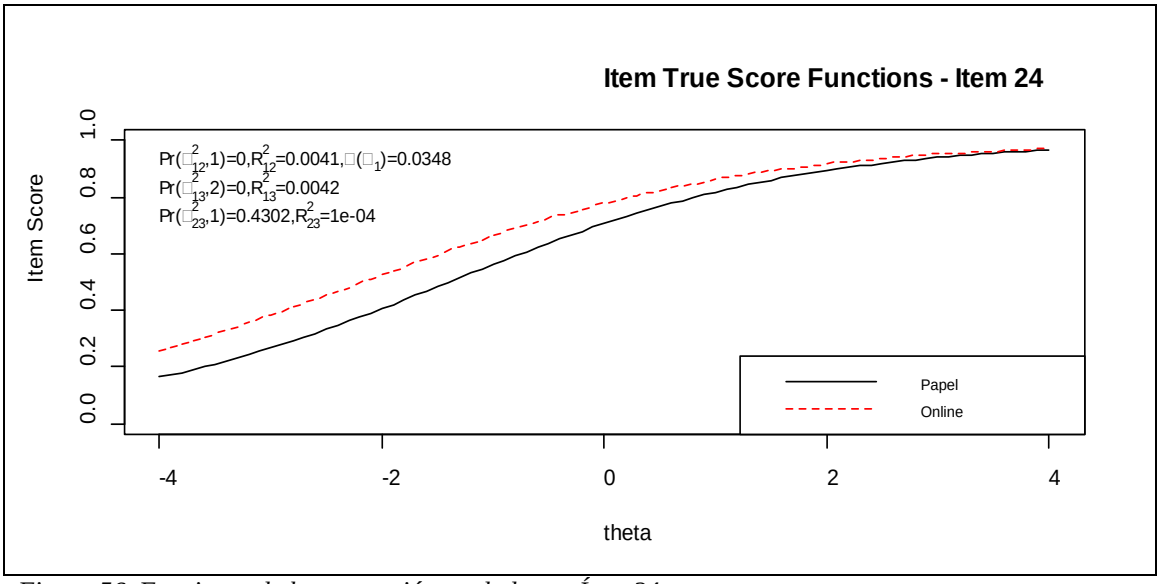


Figura 56. Funciones de la puntuación verdadera – Ítem 24

En la figura 57, podemos apreciar que las diferencias existentes se producen en los sujetos con habilidades inferiores.

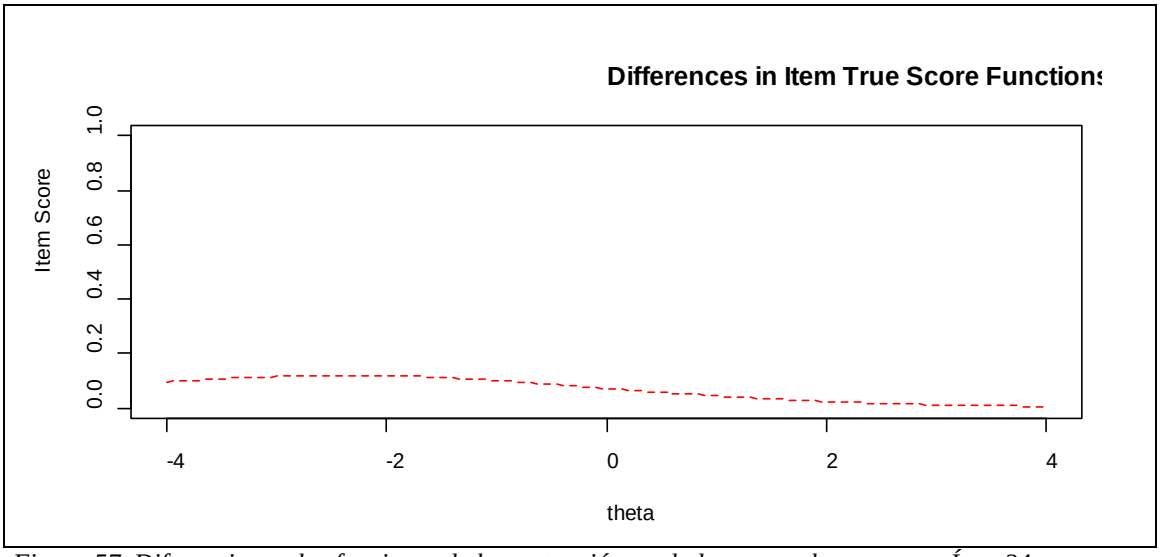


Figura 57. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 24

Por último, tal y como muestra la figura 58, la medida del efecto es muy pequeño: (R^2 < 0,035): (R^2_{12}: 0,0041; R^2_{13}: 0,0042; R^2_{23}: 0,0001), nos demuestra que el ítem 24 presenta DVF irrelevante.

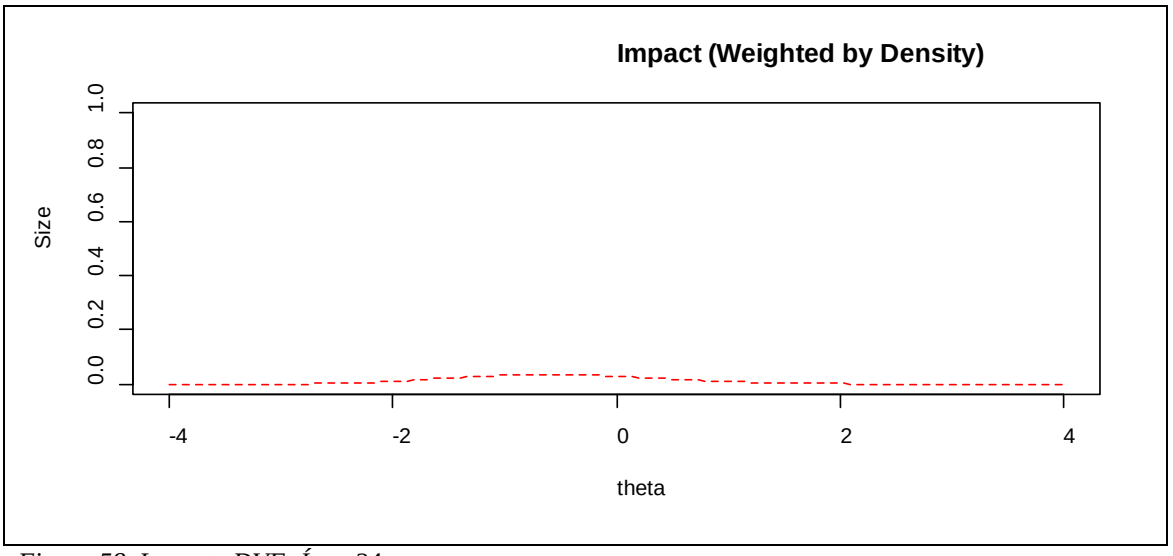


Figura 58. Impacto DVF- Ítem 24

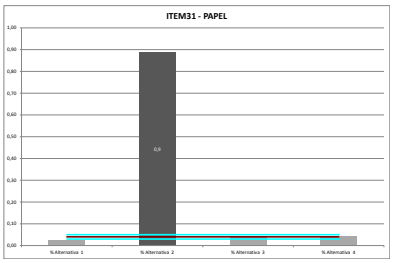
A continuación, presentamos las características relativas al ítem 31 con DVF, recogidas en la tabla 7.39.

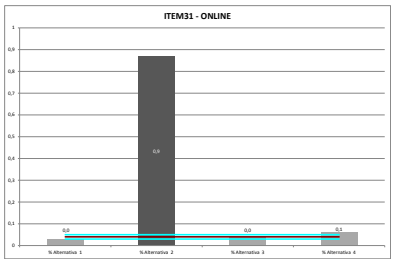
Tabla 7.39.
Características y Funcionamiento Diferencial de Versiones en el Ítem 31

Descripción desde la TCT

Ítem 31	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,89	0,10	0,32	0,36	0,30	2	23,0	17,0	0,87	0,338
Online	0,87	0,11	0,33	0,40	0,35	2	21,3	14,5	0,87	0,333

Porcentaje de elección de cada alternativa – ítem 31 en papel y online





Modelo TRI de 2 Parámetros

Ítem 31	Parámetro a	Parámetro b	p
Papel	1,254	-1972	0,655
Online	1,454	-1,769	0,714

Técnicas detección DVF Ítem 31

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,000	0,000	0,098	Uniforme	0,004	0,0045	0,0005	Débil					

*Diferencias significativas ($p \leq 0{,}0091$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

Atendiendo a las características psicométricas del ítem 31, podemos observar según la TCT que este ítem es muy fácil y discrimina poco. Concretamente la media en ambos grupos es la misma. Además los valores de los parámetros estimados según un modelo de dos parámetros nos indican valores semejantes.

En la figura 59 se recogen las funciones de respuesta al ítem para ambos grupos atendiendo a las estimaciones de los parámetros.

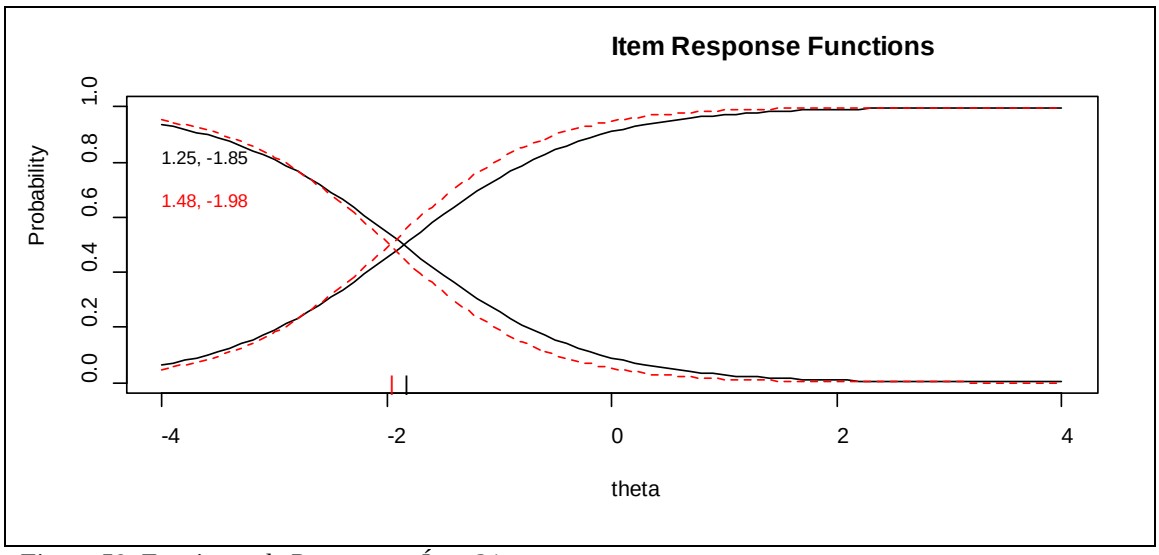


Figura 59. Funciones de Respuesta - Ítem 31

Los métodos de detección del DVF, Regresión Logística, Rajú, Lord y MH coinciden en detectar DVF en este ítem, mientras que, al igual que en casos anteriores, los métodos T.I.D. y Standard no consideran que el ítem 31 presente DVF.

En la figura 60 se recoge la significatividad (p<0,000) de la prueba $\chi 2$ para la diferencia en los modelos 1 y 3, lo que nos permite identificar este ítem como con DVF y más específicamente DVF uniforme, puesto que la prueba $\chi 2$ para la diferencia en los modelos 2 y 3 no indica diferencias significativas, p>0,01.

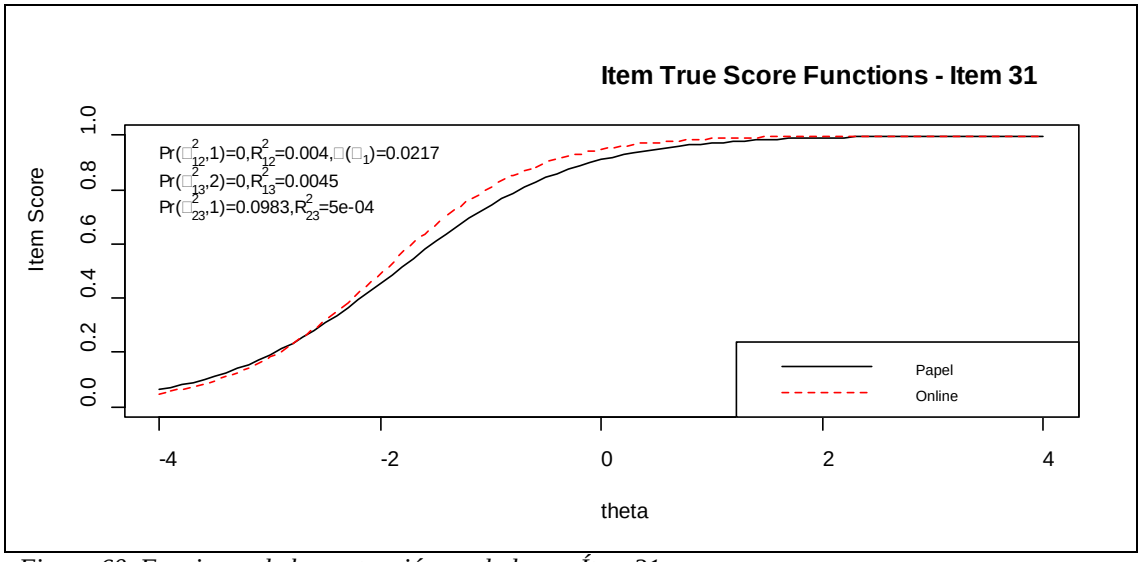


Figura 60. Funciones de la puntuación verdadera – Ítem 31

En la figura 61, podemos observar que las diferencias son más notables en los sujetos con habilidades inferiores.

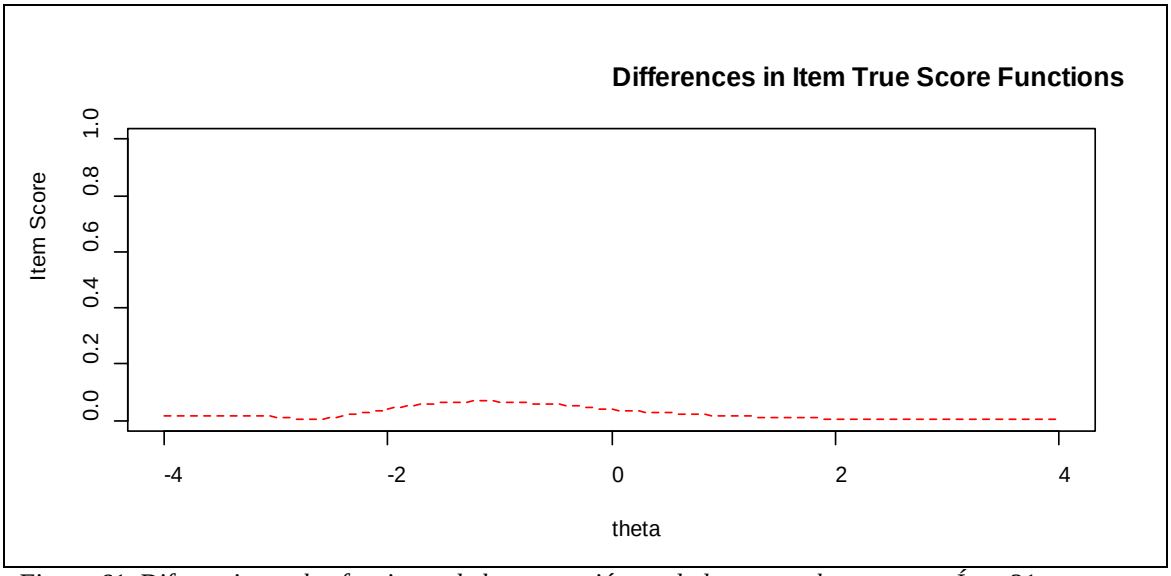


Figura 61. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 31

La figura 62, nos da evidencias del tamaño del efecto del DVF. Concretamente podemos observar que es muy pequeño. Si atendemos además a los datos podemos apreciar que el impacto es mínimo: (R^2 < 0,035): (R^2_{12}: 0,004; R^2_{13}: 0,0045; R^2_{23}: 0,0005).

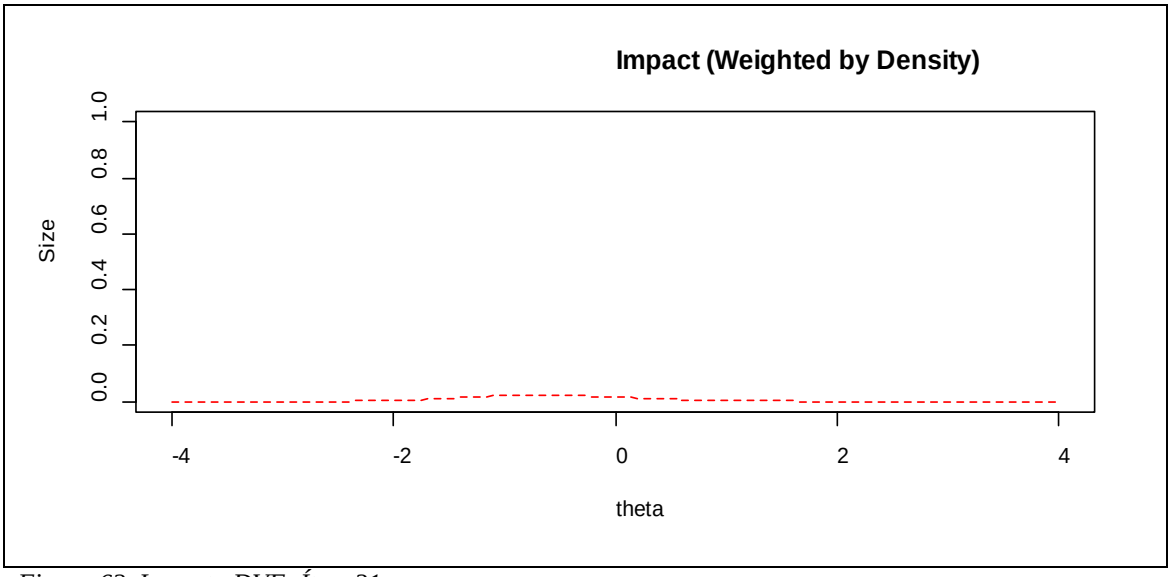


Figura 62. Impacto DVF- Ítem 31

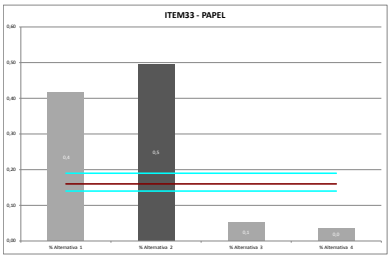
En las siguientes líneas, realizamos un análisis en detalle del ítem 13 que presenta DVF. En la tabla 7.40, se recogen las características fundamentales del ítem.

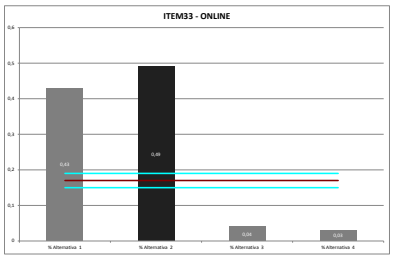
Tabla 7.40.
Características y Funcionamiento Diferencial de Versiones en el Ítem 33

Descripción desde la TCT

Ítem 33	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,50	0,25	0,50	0,43	0,35	2	24,6	20,0	0,51	0,500
Online	0,49	0,25	0,50	0,47	0,39	2	23,1	17,8	0,49	0,500

Porcentaje de elección de cada alternativa – ítem 33 en papel y online





Modelo TRI de 2 Parámetros

Ítem 33	Parámetro a	Parámetro b	p
Papel	0,980	-0,008	0,015
Online	1,084	0,031	0,125

Técnicas detección DVF Ítem 33

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DIF	No DIF	No DIF	DIF	DIF	DIF
0,000	0,000	0,187	Uniforme	0,0015	0,0017	0,0002	Débil					

*Diferencias significativas ($p \leq 0{,}0097$) nivel crítico corregido por Benjamini y HochberG

-Fuente: Elaboración propia

Las características psicométricas del ítem 33, nos indican que se trata de un ítem de dificultad media y de baja discriminación. La media en ambos grupos es bastante semejante, pero ligeramente superior en la prueba en papel.

El modelo de dos parámetros de TRI, nos indica que en la prueba online el ítem es más difícil. En la figura 63 se recogen las funciones de respuesta al ítem para ambos grupos atendiendo a las estimaciones de los parámetros.

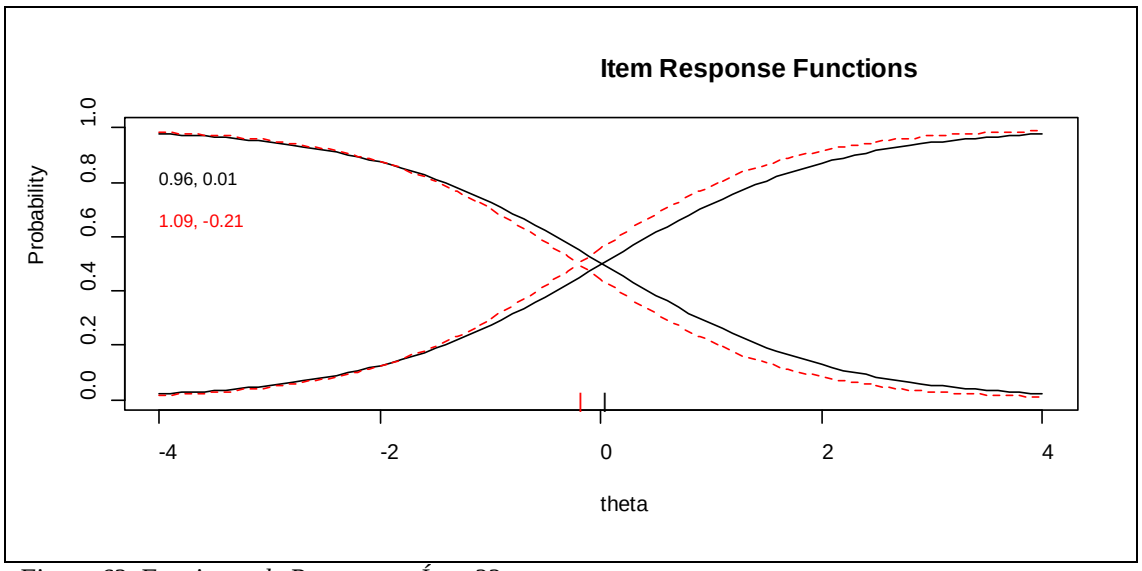


Figura 63. Funciones de Respuesta - Ítem 33

En lo que respecta al apartado de detección del DVF, además del método de Regresión Logística, los métodos Rajú, Lord y MH, también detectan este ítem con DVF. Mientras que los métodos T.I.D. y Standard no consideran que el ítem 13 presente DVF.

En la figura 64 se aprecia que el ítem 33 presenta DVF, concretamente uniforme. Los valores reflejan estos aspectos: significatividad (p<0,000) de la prueba $\chi 2$ para la diferencia en los modelos 1 y 3 y la significatividad (p>0,01) de la prueba $\chi 2$ para la diferencia en los modelos 2 y 3.

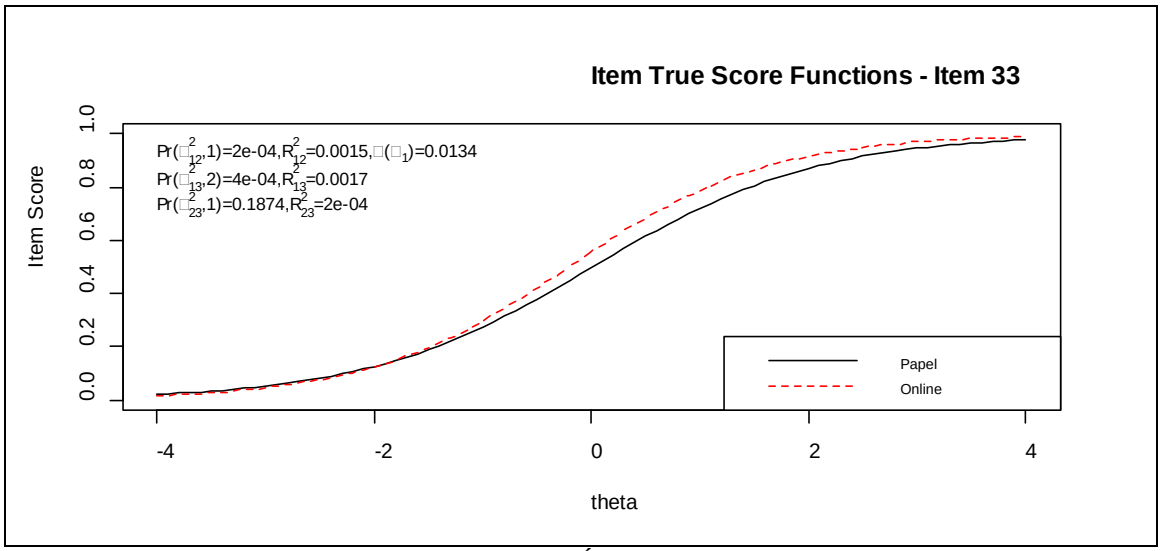


Figura 64. Funciones de la puntuación verdadera – Ítem 33

Las diferencias que se aprecian en este ítem principalmente se dan en sujetos con habilidades medias y altas (figura 65).

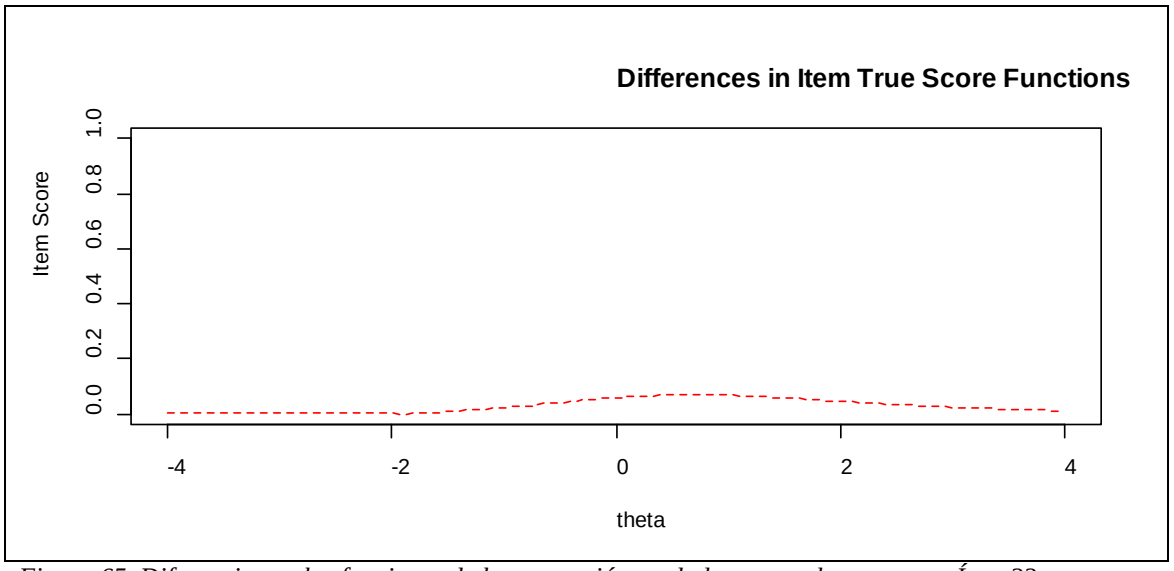


Figura 65. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 33

A pesar de estos resultados, y según se aprecia en la figura 66, la medida del efecto es muy pequeño, por lo que el impacto es mínimo: (R^2 < 0,035): (R^2_{12}: 0,0015; R^2_{13}: 0,0017; R^2_{23}: 0,0002), por ello y al igual que en los casos anteriores, el ítem 33 presenta DVF irrelevante.

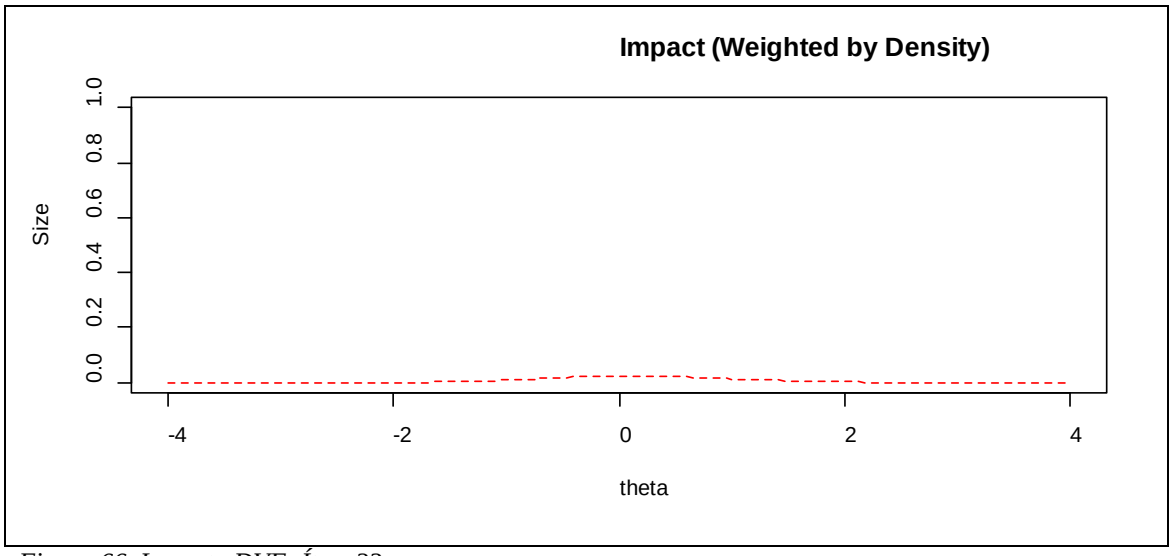


Figura 66. Impacto DVF- Ítem 33

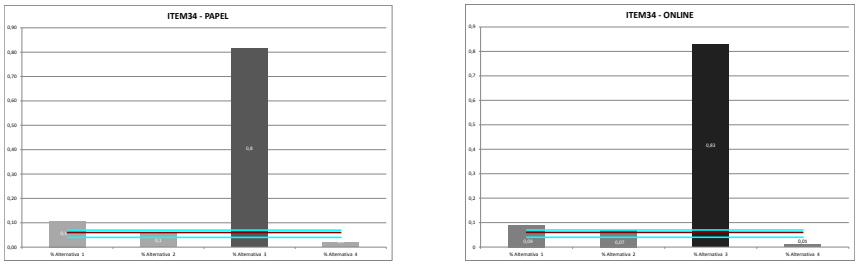
Por último, presentamos en la tabla 7.41, las características del ítem 34 que presenta DVF.

Tabla 7.41.
Características y Funcionamiento Diferencial de Versiones en el Ítem 34

Descripción desde la TCT

Ítem 34	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,81	0,15	0,39	0,33	0,27	3	23,2	18,6	0,79	0,404
Online	0,83	0,14	0,38	0,38	0,32	3	21,4	15,7	0,83	0,378

Porcentaje de elección de cada alternativa – ítem 34 en papel y online



Modelo TRI de 2 Parámetros

Ítem 34	Parámetro a	Parámetro b	p
Papel	0,933	-1,722	0,114
Online	1,183	-1,658	0,707

Técnicas detección DVF Ítem 34

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DIF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,004	No Uniforme	0,0059	0,0072	0,0013	Débil					

*Diferencias significativas (p≤0,001) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

El ítem 34 de nuevo es un ítem muy fácil que favorece algo más a los sujetos que realizan la prueba online.

En lo que respecta al modelo de TRI de dos parámetros podemos observar, en la misma línea que los resultados obtenidos en la TCT, que el parámetro de dificultad es superior en la prueba en papel, mientras que el índice de discriminación es superior en la prueba en online. Podemos ver en la figura 67 las funciones de respuesta al ítem para ambos grupos atendiendo a las estimaciones de los parámetros.

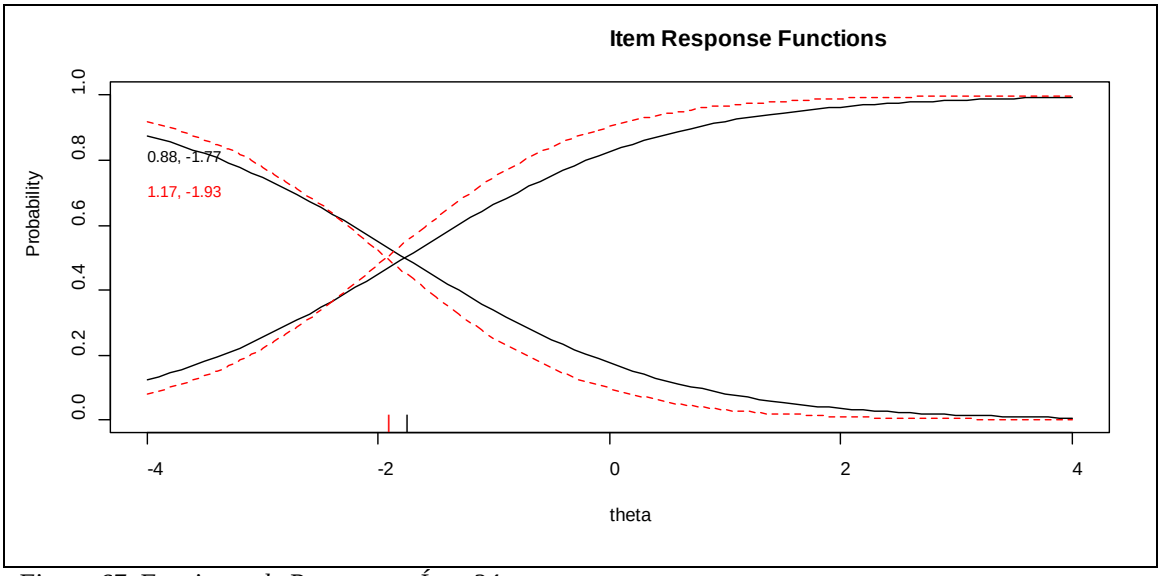


Figura 67. Funciones de Respuesta - Ítem 34

Los métodos de estimación T.I.D. y Standard no consideran que el ítem 34 presente DVF. Pero los métodos de Regresión Logística, Rajú, Lord y MH en cambio, sí detectan el ítem con DVF.

La siguiente figura 68 muestra gráficamente la situación del ítem, además viene acompañado de datos que nos permiten saber que el ítem presenta DVF (la significatividad (p<0,000) de la prueba $\chi 2$ para la diferencia en los modelos 1 y 3). Y que se trata de un ítem con DVF no uniforme (la prueba $\chi 2$ para la diferencia en los modelos 2 y 3, muestra diferencias significativas, p<0,01).

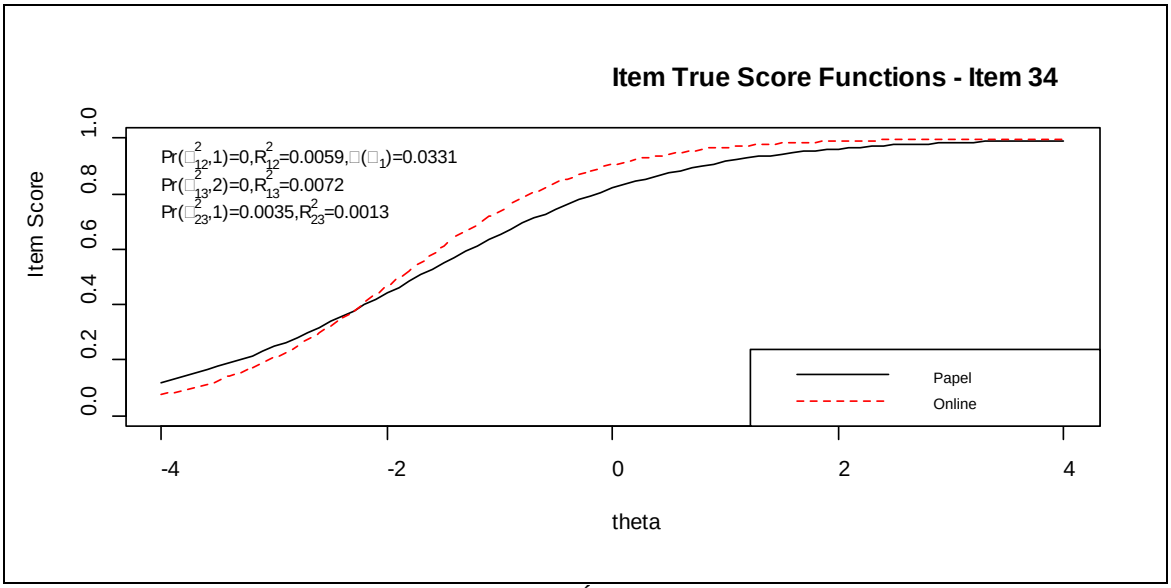


Figura 68. Funciones de la puntuación verdadera – Ítem 34

Las diferencias son más notorias en los sujetos con habilidades medias y bajas, como se refleja en la figura 69.

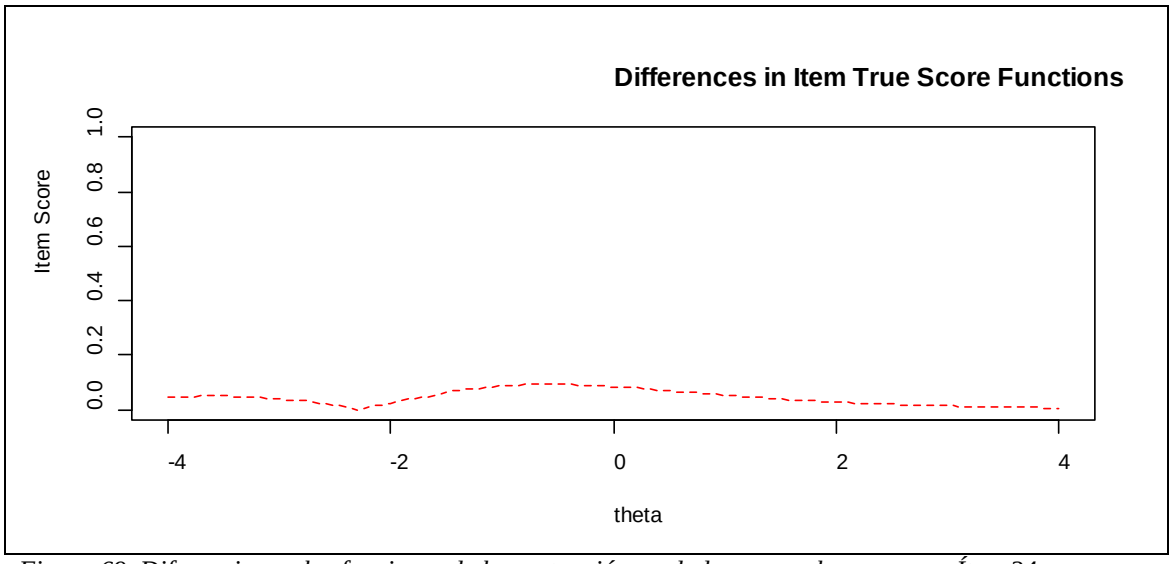


Figura 69. Diferencias en las funciones de la puntuación verdadera en ambos grupos - Ítem 34

Los resultados a pesar de lo señalado anteriormente, nos informa que la medida del efecto es muy pequeño: (R^2 < 0,035): (R^2_{12}: 0,0059; R^2_{13}: 0,0072; R^2_{23}: 0,0013), por lo que podemos considerar que el ítem 34 como todos los señalados, presenta DVF pero muy débil (ver figura 70).

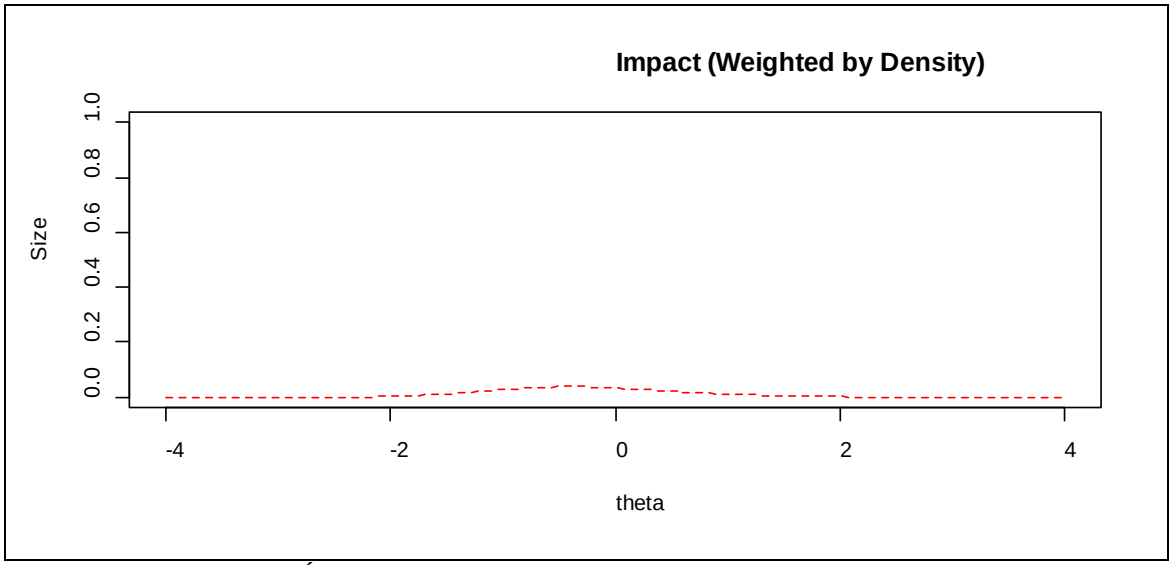


Figura 70. Impacto DVF- Ítem 34

Examinados en profundidad todos los ítems, en la tabla 7.42, puede verse a modo de resumen la descripción de cada uno de los ítems que presentan DVF.

Tabla 7.42.
Resumen ítems con Funcionamiento Diferencial de Versiones en Primaria

Ítem	Tipo DVF		Efecto DVF	T.I.D.	Stand.	Rajú	Lord	M-H
	Tipo	Grupo favorable						
4	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
7	No uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
11	No uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	DVF
13	No uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	DVF
15	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
19	No uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	No DVF
20	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
21	Uniforme	Online	DVF débil	No DVF	No DVF	DVF	DVF	DVF
23	No uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
24	Uniforme	Online	DVF débil	No DVF	No DVF	DVF	DVF	DVF
31	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
33	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
34	No uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF

Fuente: Elaboración propia

Llevado a cabo el estudio del DVF, podemos concluir de forma global, que existen ítems que presentan DVF (normalmente favorables a la prueba en papel). Pero la clase de DVF con el que nos encontramos, es débil e irrelevante, lo que nos lleva a concluir que no existe DVF; y se aportan evidencias a favor de la existencia de equivalencia entre ambas versiones de la prueba de rendimiento en comprensión lectora.

Doble contraste en Primaria

Para confirmar estos resultados y con la intención de alcanzar mayor fiabilidad en las conclusiones, realizaremos de nuevo el estudio del Funcionamiento Diferencial de Versiones utilizando otra muestra, obtenida por medio del procedimiento Propensity Score.

A la muestra escogida, se le ha aplicado la técnica Genética. En la tabla 7.43 podemos observar las características de dicha muestra.

Tabla 7.43.
Resumen de los datos de Primaria por centros, estudiantes y áreas territoriales en la prueba de Comprensión Lectora según muestra Genética

	Comprensión Lectora					
4º E.P	**Público**		**Privado Concertado**		**Privado**	
	Papel	Online	Papel	Online	Papel	Online
Centro	111	117	126	131	115	134
Norte	56	59	59	66	59	64
Sur	166	137	137	172	169	167
Este	0	0	0	0	0	0
Oeste	0	0	0	0	0	0
Total	333	345	349	369	311	365

Fuente: Elaboración propia

Conocida la muestra, a continuación se replican los análisis para el estudio del Funcionamiento Diferencial de Versiones, utilizando de nuevo la técnica de Regresión Logística.

Los resultados podemos observarlos en la tabla 7.44, donde apreciamos exactamente los mismos resultados presentados anteriormente, la presencia de DVF en los ítems: 4, 7, 11, 13, 19, 20, 21, 23, 24, 31, 33, 34. A excepción del ítem 15, que tras la corrección Benjamini y Hochberg, no presenta DVF.

A pesar de obtener presencia de DVF, al igual que en casos anteriores, el tamaño del efecto nos permite concluir que estamos de nuevo ante la presencia de DVF irrelevante.

Tabla 7.44.
Regresión Logística y Funcionamiento Diferencial de Versiones en Primaria (doble contraste)

Ítem	chi12	chi13	chi23	p-valor ajustado	Tipo de DVF	Diferencia R^2 Modelo 1 – 2	Diferencia R^2 Modelo 1 – 3	Diferencia R^2 Modelo 2 – 3	Efecto DVF
1	0,430	0,732	0,984	0,0003	No DVF	-	-	-	-
2	0,374	0,206	0,124	0,0006	No DVF	-	-	-	-
3	0,582	0,842	0,842	0,0009	No DVF	-	-	-	-
4	**0,000**	**0,000**	0,489	**0,0012**	DVF uniforme	0,0074	0,0075	0,0002	débil
5	0,556	0,837	0,921	0,0015	No DVF	-	-	-	-
6	0,719	0,394	0,188	0,0018	No DVF	-	-	-	-
7	0,006	**0,001**	**0,006**	**0,0021**	DVF no uniforme	0,0041	0,0082	0,0041	débil
8	0,469	0,668	0,595	0,0024	No DVF	-	-	-	-
9	0,033	0,030	0,119	0,0026	No DVF	-	-	-	-
10	0,134	0,324	0,929	0,0029	No DVF	-	-	-	-
11	0,004	**0,000**	**0,000**	**0,0032**	DVF no uniforme	0,0033	0,0115	0,0081	débil
12	0,422	0,653	0,648	0,0035	No DVF	-	-	-	-
13	0,004	**0,000**	**0,001**	**0,0038**	DVF no uniforme	0,0031	0,0071	0,0039	débil
14	0,257	0,508	0,794	0,0041	No DVF	-	-	-	-
15	0,005	0,019	0,903	0,0044	No DVF	-	-	-	-
16	0,253	0,248	0,224	0,0047	No DVF	-	-	-	-
17	0,657	0,725	0,504	0,0050	No DVF	-	-	-	-
18	0,870	0,650	0,361	0,0053	No DVF	-	-	-	-
19	**0,006**	0,023	0,946	0,0056	DVF no uniforme	0,0027	0,0027	0,0000	débil
20	0,127	**0,001**	**0,001**	**0,0059**	DVF uniforme	0,0009	0,0054	0,0045	débil
21	0,001	**0,000**	**0,030**	**0,0062**	DVF uniforme	0,0039	0,0055	0,0016	débil
22	0,659	0,718	0,493	0,0065	No DVF	-	-	-	-
23	0,000	**0,000**	**0,000**	**0,0068**	DVF no uniforme	0,0084	0,0282	0,0198	débil
24	**0,000**	**0,000**	0,737	**0,0071**	DVF uniforme	0,0108	0,0108	0,0000	débil
25	0,244	0,114	0,084	0,0074	No DVF	-	-	-	-
26	0,202	0,110	0,095	0,0076	No DVF	-	-	-	-
27	0,262	0,197	0,158	0,0079	No DVF	-	-	-	-
28	0,858	0,546	0,278	0,0082	No DVF	-	-	-	-
29	0,372	0,486	0,422	0,0085	No DVF	-	-	-	-
30	0,426	0,375	0,249	0,0088	No DVF	-	-	-	-
31	**0,000**	**0,001**	0,891	**0,0091**	DVF uniforme	0,0094	0,0094	0,0000	débil
32	0,966	0,445	0,203	0,0094	No DVF	-	-	-	-
33	**0,000**	**0,001**	0,251	**0,0097**	DVF uniforme	0,0044	0,0049	0,0005	débil
34	**0,000**	**0,000**	0,184	**0,0100**	DVF no uniforme	0,0177	0,0185	0,0009	débil

Fuente: Elaboración propia
Nota: En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 3 significativas (atendiendo al p- valor ajustado). Lo que significa la presencia de DVF.
En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 2 significativa (p- valor ajustado) (presenta DVF uniforme)
En negrita se indican las diferencias significativas entre el Modelo 2 - Modelo 3 significativa (p- valor ajustado) (presenta DVF no uniforme)
Diferencia R^2_{12} , R^2_{13} , R^2_{23} ($p<0,035$) (DVF débil o insignificante)

7.6.2.2. Descripción de los ítems con Funcionamiento Diferencial de Versiones en Secundaria

Al igual que llevamos a cabo el estudio del DVF en Primaria, se ha replicado para Secundaria. Con la intención de no ser reiterativos, se adjuntan en el anexo 37 las tablas con las características de cada ítem que presenta DVF, así como todas las gráficas que el paquete Lordif ha originado. A modo de resumen, en la tabla 7.45, son presentados los resultados de Secundaria.

Tabla 7.45.
Resumen ítems con Funcionamiento Diferencial de Versiones en Secundaria

Ítem	Tipo DVF		Efecto DVF	T.I.D.	Stand.	Rajú	Lord	M-H
	Tipo	Grupo favorable						
1	Uniforme	Online	DVF débil	No DVF	No DVF	No DVF	DVF	DVF
2	Uniforme	Online	DVF débil	No DVF	No DVF	No DVF	DVF	DVF
3	Uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	DVF
5	Uniforme	Online	DVF débil	No DVF	No DVF	No DVF	DVF	DVF
9	Uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	DVF
11	No uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	DVF
12	No uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
13	No uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	No DVF
14	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
15	No uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
16	Uniforme	Online	DVF débil	No DVF	No DVF	DVF	DVF	DVF
18	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
20	Uniforme	Online	DVF débil	No DVF	DVF	DVF	DVF	DVF
21	No uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
22	Uniforme	Papel	DVF débil	No DVF	No DVF	DVF	DVF	DVF
24	Uniforme	Online	DVF débil	No DVF	No DVF	DVF	DVF	DVF
25	Uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	No DVF
26	Uniforme	Online	DVF débil	No DVF	No DVF	DVF	DVF	DVF
30	Uniforme	Online	DVF débil	No DVF	No DVF	DVF	DVF	DVF
33	No uniforme	Papel	DVF débil	No DVF	No DVF	No DVF	No DVF	No DVF

Fuente: Elaboración propia

Tras el estudio del DVF en Secundaria, podemos hablar de la existencia de ítems que efectivamente presentan DVF, favorables en algunos casos a la versión en papel y en otros casos a la versión online. Pero de nuevo, y al igual que en Primaria, nos encontramos con DVF irrelevante. Por lo tanto, volvemos a confirmar la similitud

esperable y concluimos que el modo de aplicación de las versiones no provoca DVF y son equivalentes.

Doble contraste en Secundaria

Al igual que en Primaria, llevamos a cabo un doble contraste con otra muestra para verificar los resultados obtenidos. Por ello, realizamos nuevamente el Funcionamiento Diferencial de Versiones con la muestra obtenida con el procedimiento Propensity Score y la técnica Genética.

Las características de la muestra son recogidas en la tabla 7.46.

Tabla 7.46.
Resumen de los datos de Secundaria por centros, estudiantes y áreas territoriales en la prueba de Comprensión Lectora según muestra Genética

	Comprensión Lectora					
2º E.S.O	**Público**		**Privado Concertado**		**Privado**	
	Papel	Online	Papel	Online	Papel	Online
Centro	254	256	718	738	341	381
Norte	199	212	129	142	122	133
Sur	338	345	0	0	0	0
Este	0	0	0	0	0	0
Oeste	0	0	0	0	0	0
Total	731	813	847	880	463	514

Fuente: Elaboración propia

En la tabla 7.47, podemos observar los ítems que presentan DVF en la nueva muestra; concretamente, se tratan de 14 ítems: 5, 6, 12, 14, 15, 16, 18, 20, 21, 22, 24, 30, 32, 33 (son 6 ítems menos que en la muestra anterior de Secundaria)

Tabla 7.47.
Regresión Logística y Funcionamiento Diferencial de Versiones en Secundaria (doble contraste)

Ítem	chi12	chi13	chi23	p-valor ajustado	Tipo de DVF	Diferencia R^2 Modelo 1 – 2	Diferencia R^2 Modelo 1 – 3	Diferencia R^2 Modelo 2 – 3	Efecto DVF
1	0,014	0,045	0,729	0,0003	No DVF	-	-	-	-
2	0,418	0,390	0,268	0,0006	No DVF	-	-	-	-
3	0,264	0,175	0,134	0,0009	No DVF	-	-	-	-
4	0,878	0,197	0,073	0,0012	No DVF	-	-	-	-
5	**0,000**	**0,000**	0,257	**0,0015**	DVF uniforme	0,0067	0,0071	0,0004	débil
6	0,026	**0,000**	**0,000**	**0,0018**	DVF no uniforme	0,0010	0,0035	0,0025	débil
7	0,806	0,897	0,693	0,0021	No DVF	-	-	-	-
8	0,004	0,006	0,165	0,0024	No DVF	-	-	-	-
9	0,106	0,190	0,399	0,0027	No DVF	-	-	-	-
10	0,395	0,614	0,614	0,0030	No DVF	-	-	-	-
11	0,021	0,012	0,063	0,0033	No DVF	-	-	-	-
12	**0,000**	**0,000**	0,001	**0,0036**	DVF uniforme	0,0085	0,0108	0,0023	débil
13	0,428	0,344	0,220	0,0039	No DVF	-	-	-	-
14	**0,000**	**0,000**	0,061	**0,0042**	DVF uniforme	0,0030	0,0036	0,0006	débil
15	**0,000**	**0,000**	0,600	**0,0045**	DVF uniforme	0,0046	0,0047	0,0000	débil
16	**0,000**	**0,001**	0,844	**0,0048**	DVF uniforme	0,0025	0,0025	0,0000	débil
17	0,963	0,449	0,206	0,0052	No DVF	-	-	-	-
18	**0,000**	**0,000**	0,745	**0,0055**	DVF uniforme	0,0042	0,0043	0,0000	débil
19	0,279	0,479	0,585	0,0058	No DVF	-	-	-	-
20	**0,000**	**0,000**	0,754	**0,0061**	DVF uniforme	0,0166	0,0166	0,0000	débil
21	**0,001**	**0,000**	0,000	**0,0064**	DVF uniforme	0,0025	0,0069	0,0043	débil
22	**0,002**	**0,007**	0,409	**0,0067**	DVF uniforme	0,0016	0,0017	0,0001	débil
23	0,440	0,733	0,872	0,0070	No DVF	-	-	-	-
24	**0,000**	**0,000**	0,729	**0,0073**	DVF uniforme	0,0083	0,0083	0,0000	débil
25	0,049	0,118	0,526	0,0076	No DVF	-	-	-	-
26	0,169	0,044	0,037	0,0079	No DVF	-	-	-	-
27	0,885	0,316	0,131	0,0082	No DVF	-	-	-	-
28	0,035	0,011	0,031	0,0085	No DVF	-	-	-	-
29	0,697	0,921	0,911	0,0088	No DVF	-	-	-	-
30	**0,000**	**0,000**	0,013	**0,0091**	DVF uniforme	0,0086	0,0109	0,0023	débil
31	0,109	0,127	0,212	0,0094	No DVF	-	-	-	-
32	**0,011**	**0,003**	0,022	**0,0097**	DVF uniforme	0,0013	0,0023	0,0010	débil
33	0,098	**0,001**	**0,001**	**0,0100**	DVF no uniforme	0,0005	0,0027	0,0022	débil

Fuente: Elaboración propia
Nota: En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 3 significativas (atendiendo al p- valor ajustado). Lo que significa la presencia de DVF.
En negrita se indican las diferencias significativas entre el Modelo 1 - Modelo 2 significativa (p- valor ajustado) (presenta DVF uniforme)
En negrita se indican las diferencias significativas entre el Modelo 2 - Modelo 3 significativa (p- valor ajustado) (presenta DVF no uniforme)
Diferencia R^2_{12} , R^2_{13} , R^2_{23} ($p<0,035$) (DVF débil o insignificante)

Los resultados alcanzados arrojan las mismas conclusiones, a pesar de detectar presencia de DVF en estos ítems, el tamaño del efecto es mínimo, por tanto estamos ante la presencia de ítems con DVF irrelevante o débil.

CAPÍTULO 8: Discusión y conclusiones

"Ciencia es todo aquello sobre lo cual siempre cabe discusión"

José Ortega y Gasset

En las siguientes líneas se presentan las principales conclusiones sobre la equivalencia de las pruebas de rendimiento en Comprensión Lectora atendiendo al modo de aplicación (convencional e informatizada). Posteriormente se presentan las limitaciones y futuras líneas de investigación tras la realización de este trabajo.

Una de las aportaciones más relevantes, en torno a la cual gira esta investigación, es la propuesta metodológica llevada a cabo para estudiar el posible sesgo que se puede ver reflejado en el rendimiento y puntuaciones de los evaluados, provocado por el modo en que se aplica la prueba (test en papel y lápiz o test informatizados). Es por ello, que ha sido necesario el estudio del primero de los objetivos planteados: "*Demostrar la utilidad de las Propensity Score para el emparejamiento efectivo de muestras y para el estudio de la equivalencia y de la detección del Funcionamiento Diferencial de Versiones en ambas versiones de una prueba de rendimiento en comprensión lectora"*.

Tras el empleo de las técnicas de Propensity Score se han logrado grupos homogéneos que nos permiten realizar correctamente los estudios del DVF. Concretamente, estábamos ante una muestra desequilibrada: Primaria (N_{papel} = 9.258; $N_{online\ l}$= 1.079); Secundaria (N_{papel} =46.482; N_{online} = 2.207). Tras el balance realizado, la muestra alcanzada es más homogénea, tanto para Primaria (N_{papel} = 5.486; N_{online} = 1.079) como para Secundaria (N_{papel} = 14.511; N_{online} = 2.207).

Como ya hemos adelantado en líneas anteriores, esta técnica nos permite reducir el sesgo de selección y garantizar que las diferencias que nos encontremos se deban al modo de aplicación y no a otras variables. Además, se logra que sujetos con las mismas características tengan una puntuación tanto en la prueba en papel como en la prueba online, lo que hace posible valerse de las Propensity Score para el estudio del DVF.

Una vez disponibles las muestras equivalentes se procede a dar respuesta a los dos problemas de investigación planteados, atendiendo a los objetivos e hipótesis propuestas.

Problema de Investigación 1

"¿El modo en que es aplicado un test (papel y online) provoca diferencias estadísticamente significativas en el rendimiento en comprensión lectora?"

El estudio de la equivalencia de ambas versiones, primer objetivo planteado, nos permite concluir que existe equivalencia entre ambas versiones. Son numerosos los estudios en los que el modo de aplicación del test no produce diferencias estadísticamente significativas, como el caso de Sawaki (2001) que recoge el estudio de Spray, Ackerman, Reckase y Carlson en 1989, donde las puntuaciones medias, así como sus distribuciones, no presentan diferencias significativas en función del modo de presentación de los test.

En lo que concierne al test de rendimiento, una de las primeras revisiones sobre la comparación de estas pruebas fue llevada a cabo por Mazzeo y Harvey (1998), quienes compararon los resultados de treinta investigaciones de todo tipo (personalidad, aptitud, inteligencia). Brand y Houx (1992) y Rolls y Feltham (1993), en sus estudios con test de personalidad y de actitudes, señalan que tampoco se producen diferencias significativas en función del modo de aplicación de la prueba.

Čandrlić, Ašenbrener y Holenko (2014) recientemente llevaron a cabo una investigación con la intención de comprobar la equivalencia entre las versiones convencionales y las versiones online; para ello aplicaron, en tres cursos diferentes, pruebas en asignaturas del Departamento de Informática de la Universidad de Rijeka (Croacia). Los resultados demostraron que no existen diferencias estadísticamente significativas entre ambas versiones.

A la misma conclusión llegaron Eleanor, Ashley y Morin (2014) aplicando ambas versiones de una prueba en estudiantes de ingeniería; a pesar de que los resultados desvelaron que no existen diferencias estadísticamente significativas entre el resultado obtenido en papel y en el ordenador, el 86% de los estudiantes prefirieron realizar la prueba en papel.

La conclusión a la que hemos llegado ha sido determinada por el análisis descriptivo y psicométrico llevado a cabo, tanto en Primaria como en Secundaria. Mediante este análisis hemos comprobado que las dispersiones, distribuciones de las puntuaciones, las medias (Primaria: $\bar{x}_{papel}$=22,22; $\bar{x}_{online}$=20,44; Secundaria: $\bar{x}_{papel}$=24,09; $\bar{x}_{papel}$=23,36), así como la fiabilidad (Primaria: α_{papel}=22,22; α_{online}=20,44; Secundaria: α_{papel}=24,09; α_{papel}=23,36) son equivalentes en ambos modos de aplicación.

También, atendiendo a la Teoría de Respuesta al Ítem, se observa que el modelo que mejor ajusta en la prueba en papel y en la prueba online coincide y corresponde con el modelo de dos parámetros, tanto en Primaria como en Secundaria.

Similares resultados se obtienen cuando se realiza el estudio de la estructura unidimensional de las versiones. La estructura factorial es invariante y prácticamente idéntica (Primaria: 23% de varianza explicada tanto en papel como online; Secundaria 27% de varianza explicada en la prueba en papel y 26% en la prueba online). Además, al realizar el análisis de invarianza factorial los resultados nos demuestran la existencia de igualdad de varianzas de los factores, igualdad de las covarianzas entre los factores e igualdad de medias en ambas versiones.

Tras el estudio de la homogeneidad de varianzas y la comparación de medias entre ambas versiones, los resultados nos indican que la puntuación es superior en los sujetos que realizan la prueba en papel, siendo más fácil que la prueba online, observándose diferencias algo más notables en Secundaria que en Primaria. A pesar de ello, el tamaño del efecto es muy pequeño en ambos casos, confirmando la similitud esperable y concluyendo la equivalencia entre ambas versiones.

Los resultados de las pruebas Electronic Reading Assessment – PISA-ERA (OCDE, 2009), que miden concretamente la lectura digital, muestran menores puntuaciones en lectura digital que en lectura impresa por parte de los estudiantes españoles (6 puntos menos). Al igual que en PISA 2012 (OCDE, 2013), donde se realizaron estudios comparativos y, de nuevo, el resultado demuestra que los estudiantes

españoles obtienen una media inferior en la versión informatizada (22 puntos menos) en la prueba de comprensión lectora.

Valorada la prueba de forma global, se han estudiado de forma particular los ítems. Los resultados nos permiten concluir que aproximadamente en la mitad de los ítems el modo de aplicación no influye en la puntuación que obtiene el sujeto.

Pero existe un conjunto de ítems que reflejan la existencia de diferencias en su puntuación en función del modo en que el ítem ha sido aplicado. A pesar de estas diferencias, es necesario mencionar que el tamaño del efecto es muy pequeño. Al igual que sucede con los estudios llevados a cabo por Karkeen, Kim y Fatica (2010), Kingston (2008) y Lottridge, Nicewander y Mitzel (2011), donde se informa de tamaños pequeños del efecto, que indican que las diferencias en las medias no eran significativas.

Si analizamos estos ítems, observamos que todos ellos tienen una media superior cuando son realizados en papel. En Primaria tan solo el ítem 24 obtiene una media superior en el modo online; en Secundaria son los ítems 1, 5, 16, 20 y 24 los que arrojan una media superior cuando se realizan online.

Podemos apreciar que la diferencia de medias y puntuaciones en estos ítems favorecen a la prueba en papel; esta idea es compartida por autores como Bennett et al. 2008; Karkeen, Kim y Fatica, 2010; Jackel, 2014; Sandene, et al. 2005; y Way, Davis y Fitzpatrick, 2006.

A la misma conclusión llegan Puhan, Boughton y Kim (2007), que consideran en su estudio que ambas versiones son equivalentes, puesto que las puntuaciones medias no eran significativamente diferentes en ambas versiones; pero estas mínimas diferencias mostraban mejores resultados en la versión en papel. También nos informan de tamaños del efecto pequeños y no significativos, que indican la no diferencia entre las dos versiones de la prueba.

Estas diferencias con tamaños del efecto pequeños puede deberse a las limitaciones que las aplicaciones online tienen, y que pueden repercutir en el resultado

obtenido, favorable a la prueba en papel. Principalmente se refieren a la ansiedad que puede ocasionar para el evaluado la realización de la prueba en ordenador por la escasa familiaridad con el medio (Goldberg y Pedulla, 2002; Pomplun, Ritchie y Custer, 2006). Olea, Ponsoda y Prieto (1999) recogen el estudio de Hedl, O´Neil y Hansen llevado a cabo en 1973, en el que observaron elevados grados de ansiedad y actitud negativa ante la aplicación informatizada del instrumento WAIS y el Slosson Intelligence Test, debido a la poca experiencia en el manejo del ordenador o a la falta de instrucciones claras. También destaca, como contraposición, el estudio de Wise, Barnes, Harvey y Plake en 1989, en el que no observaron ansiedad significativa ante la aplicación informatizada en dos grupos que realizaron una prueba de matemáticas.

Hay que tener en cuenta que la exposición que actualmente tienen los niños a los ordenadores es masiva, por lo que estos estudios pueden haber quedado anticuados. Autores como Cabras y Tena (2013) afirman que *"el posible efecto positivo parece ser significativamente mayor en estudiantes que pertenecen a grupos socioeconómicos más desfavorecidos, lo que refuerza la consideración de este tipo de intervención como una herramienta para conseguir mayor equidad"* (p.69).

Jeong (2014) lleva a cabo un estudio con estudiantes familiarizados con los ordenadores para verificar si esta característica beneficiaba a los estudiantes que realizaban la prueba online. Los resultados alcanzados reportaban mayores puntuaciones en la prueba en papel, verificando la no existencia de relación entre ambas variables. Esta idea es compartida por Marcenaro (2014).

Por otro lado, autores como Mourant, Lakshmanan y Chantadisai (1981) han estudiado la fatiga o el cansancio producidos por la lectura en la pantalla del ordenador, y los resultados han demostrado la existencia de este cansancio (Russell y Haney, 1997). Dillon (1994) concreta que la lectura por ordenador en pruebas de rendimiento era de un 20% a un 30% más lenta que en papel.

De ahí surge la importancia de que el texto en versiones informatizadas sea presentado en su totalidad en la pantalla. Russell y Haney (1997), haciéndose eco de las ideas de Haas y Hayes en 1986, observan puntuaciones más bajas en las versiones

informatizadas que en las versiones de papel y lápiz cuando el texto aparece en más de una página, fruto de la complejidad que conlleva una lectura de este tipo. Muchos estudios han demostrado que si un ítem puede ser presentado en su totalidad en la pantalla del ordenador, las diferencias según el modo de administración de la prueba son nulas o casi inexistentes (TEA, 2008).

Cuando el texto no aparece en su totalidad en la pantalla, existen diferencias significativas, en general a favor de los estudiantes que realizan el test en papel y lápiz (Bergstrom, 1992; Bridgeman et al., 2001; Higgins et al., 2005; Keng et al., 2008; O'Malley et al., 2005, citado en TEA, 2008; Pommerich, 2004). Mientras que Belmore (1985) (citado en Noyes y Garland, 2008); McGoldrick et al., 1992 (citado en Sawaki, 2001) y Zuk (1986, citado en Paek, 2005) destacan que la lectura se realizaba en menor tiempo en pantalla que en papel.

Otros inconvenientes son los relativos a las características de los ordenadores, concretamente a aspectos como que el tamaño de las pantallas no sea siempre el mismo, que las condiciones no siempre sean las deseadas (resolución, tamaño de la fuente) o problemas con el software de navegación y configuración, todo lo cual puede producir diferencias en los resultados (McKee y Levinson, 1990; Noyes y Garland, 2008).

Problema de Investigación 2

> ***"¿El modo en que es aplicado un test (papel y online) provoca Funcionamiento Diferencial de Versiones en pruebas de rendimiento en comprensión lectora?"***

Tras el estudio del Funcionamiento Diferencial de Versiones (DVF) podemos determinar que el modo de aplicación de un test no afecta al DVF; por tanto, podemos concluir que ambas versiones de la prueba de rendimiento en Comprensión Lectora, en Primaria así como en Secundaria, son equivalentes.

Nuestra propuesta metodológica, el "Funcionamiento Diferencial de Versiones", se aproxima a los estudios del Funcionamiento Diferencial de los Ítems y de los Sujetos, pero esta metodología necesita de la utilización de las técnicas de Puntaje de Propensión, o Propensity Score, para conseguir sujetos homogéneos que nos permitan comparar las puntuaciones obtenidas en ambas versiones. Disponibles las muestras equivalentes (objetivo 1), es posible estudiar el tercer objetivo propuesto: *"evaluar la equivalencia de una prueba de rendimiento en comprensión lectora elaborada en dos versiones (papel y lápiz y online) por medio del estudio del Funcionamiento Diferencial de Versiones.*

A la vista de las técnicas expuestas para la detección del Funcionamiento Diferencial de Versiones, el procedimiento de Regresión Logística –tras la corrección de Benjamini y Hochberg (1995) utilizando el método "*False Discovery Rate*" o Tasa de Falsos Descubrimientos– detecta en Primaria 13 ítems y en Secundaria 20 ítems con DVF, favorable principalmente al grupo de sujetos que realizan la prueba en papel. La medida del efecto del DVF nos indica que la presencia de DVF es irrelevante o muy débil en todos los casos.

Con la intención de ser más precisos en nuestras conclusiones, también se ha llevado a cabo un doble contraste utilizando otra muestra obtenida por medio del Puntaje de Propensión o Propensity Score, con la técnica "Genetic Matching" o Genética. Los resultados obtenidos vuelven a ser los mismos en Primaria (mismos ítems con DVF) y en Secundaria (14 ítems presentan DVF, 6 ítems menos que en la muestra anterior). Aun así, en ambas etapas educativas el tamaño del efecto es, de nuevo, mínimo, considerando igual que anteriormente irrelevante o débil la existencia de DVF.

Para garantizar conclusiones más consistentes, se ha atendido a otros procedimientos basados en la Teoría Clásica de los test (como el procedimiento T.I.D.), en la Teoría de Respuesta al Ítem (Rajú y Lord) y en las tablas de contingencia (Estandarizado-Stand, Mantel – Haenszel).

En Primaria, de los 13 ítems que presentan DVF, los procedimientos T.I.D. y Estandarizado no detectan estos ítems con DVF. En cambio, los procedimientos de Rajú, Lord y Mantel-Haenszel detectan tan solo 3 ítems de los 13 con DVF.

En Secundaria, según el procedimiento de Regresión Logística, contábamos con 20 ítems con DVF, y los procedimientos T.I.D no detectan ninguno de los 20 ítems con DVF; en el caso del procedimiento Estandarizado tan sólo uno de los ítems es detectado con DVF. Los procedimientos de Rajú, Lord y Mantel-Haenszel detectan 11, 14 y 17 ítems con DVF, respectivamente.

Dado que estos procedimientos no presentan medidas del efecto, se atiende a los resultados obtenidos por la regresión logística, ya que esta medida nos permite conocer la magnitud del efecto para evitar los falsos DVF (Hidalgo, Gómez-Benito y Zumbo, 2014).

Por todo ello, podemos concluir que el modo de aplicación de un test no afecta significativamente al Funcionamiento Diferencial de Versiones y, por consiguiente, podemos determinar que ambas versiones (papel y online) son equivalentes.

Estos resultados coinciden con los obtenidos en el estudio realizado por Seo y Jong (2015), donde se habla de equivalencia entre ambas versiones debido a la no existencia de DIF, ni diferencias estadísticamente significativas en el rendimiento medio de los estudiantes en función del modo de aplicación.

Resultados afines son los obtenidos por Keng, McClarty y Davis (2008) y Poggio et al. (2005), que, con el estudio del funcionamiento diferencial de los ítems y el estudio comparativo, encuentran pocas preguntas de matemáticas (9 de 204 ítems) que se comporten de manera diferente en las dos modalidades.

En cuanto al trabajo llevado a cabo por Lottridge, Nicewander y Mitzel (2011), se comparan un test aplicado en papel y otro online de Algebra e Inglés. Para realizar el

estudio comparativo utilizaron la técnica de Propensity Score, logrando que ambos grupos fueran homogéneos. Los resultados alcanzados demostraban que ambas versiones eran comparables, dado que obtenían parejos niveles de fiabilidad, correlaciones cercanas a 1 y estructuras factoriales paralelas; la dificultad era mayor en la prueba online, pero el tamaño del efecto era mínimo, por lo que las diferencias en las medias no eran significativas.

En vista de todos los resultados obtenidos, tras estudiar el test de Comprensión Lectora, y llevando a cabo un estudio comparativo y logrando grupos homogéneos por medio del Puntaje de Propensión o Propensity Score, los resultados demuestran que ambas versiones son comparables, obteniendo equivalentes distribuciones, niveles de fiabilidad correctos e iguales, ajuste adecuado en el modelo de dos parámetros, estructuras factoriales paralelas y la no existencia de Funcionamiento Diferencial de Versiones. La media tiende a favorecer a la prueba en papel, considerándose ésta algo más fácil que la prueba online, pero el tamaño del efecto es mínimo, por lo que podemos concluir que ambas versiones son equivalentes.

Estos resultados muestran la precisión tenida en cuenta en el diseño y la elaboración de los instrumentos de Comprensión Lectora. El logro de la adaptación y equivalencia de ambas versiones ha sido posible porque se ha atendido a los *"Standards for Educational and Psychological Testing"* (AERA, APA, y NCME, 2014) y a las directrices marcadas para las pruebas aplicadas en ordenador/internet: "International Guidelines on Computer-Based and Internet Delivered Testing" (International Test Commission, 2005).

Verificada esta equivalencia y la no existencia de Funcionamiento Diferencial de Versiones en las pruebas de Diagnóstico de la Comunidad de Madrid, podríamos favorecernos de los muchos beneficios que proporcionan las versiones informatizadas:

El ordenador es la vía perfecta para el almacenamiento e introducción automática de datos, lo que facilita la obtención de resultados e informes.

La llegada del ordenador al ámbito evaluativo supone un cambio en la edición de los ítems (ya que abre nuevos canales visuales y auditivos), permite fijar el tiempo de exposición de las preguntas (lo que conlleva mayor control, garantizando las mismas condiciones para todos), hay una menor probabilidad de copia (con la utilización de presentaciones aleatorizadas de los ítems) y, además, supone menor tiempo de aplicación, menor número de ítems y con ello menores costes (Csapó, Ainley, Bennett, Latour y Law, 2012).

En lo que atañe a la introducción de datos, Olea y Hontangas (1999), Molina, Sanmartín y Pareja (2000), y Garcés, Sepúlveda y Riquelme (2014) señalan que el ordenador evita errores, tanto por parte del investigador (tabulando los datos en el ordenador) como por parte del sujeto (cuando contesta en la hoja de respuesta, donde puede confundirse).

Asimismo, supone la obtención de la puntuación de forma automática, lo que implica la automatización de los informes y su debida interpretación de los datos (Schade, Hernández y Elgueta, 2005).

Limitaciones y futuras líneas de investigación

Por último, queremos señalar algunas limitaciones que este trabajo de investigación presenta, así como posibles líneas de investigación futuras que abordaremos a continuación.

La Evaluación de Diagnóstico se caracteriza por la evaluación de diferentes destrezas además de la Comprensión Lectora (concretamente Lengua y Matemáticas). En este trabajo únicamente se aborda la destreza de Comprensión Lectora, lo que limita el obtener una perspectiva global, además de comparada, entre todas las destrezas; por ello sería conveniente hacer un estudio parejo al presentado comparando ambas versiones en cada una de las destrezas.

En esta línea, se ha realizado un estudio para la prueba de diagnóstico en la destreza de matemáticas (Jiménez y González, 2015) con los mismos resultados

alcanzados en esta investigación, la no existencia de funcionamiento diferencial de los ítems y la equivalencia entre ambas versiones (papel y online) de la prueba; pero habría que aplicar la propuesta metodológica empleada en este trabajo para garantizar la no existencia de sesgo debido a la no selección aleatoria.

Otra limitación del presente trabajo, a pesar del rigor metodológico utilizado para garantizar resultados lo más exactos posibles, es la imposibilidad de generalizar los resultados; aunque no fuera éste el propósito de dicha investigación. Hablar de equivalencia entre versiones depende de las pruebas utilizadas en el estudio de comparación, pero recomendamos algunas directrices que debemos seguir para garantizar dicha equivalencia.

Una de las futuras líneas de investigación a este respecto sería la elaboración de Test Adaptativos Informatizados, dado que garantizan una mayor seguridad y un menor tiempo de aplicación. Muñiz y Hambleton (1999) hacen hincapié en la ventaja que proporcionan estos test en lo relativo al tiempo de aplicación, dado que la prueba se adapta al nivel de cada sujeto y se ahorra tiempo debido a que los test son más cortos, por ello, con el mismo número de ítems que en los test convencionales y en los test convencionales informatizados, podemos alcanzar estimaciones más precisas.

En lo que respecta al diseño empleado y al tamaño de los grupos estudiados, es posible la existencia de sesgo debido a que los mismos centros tomaron la decisión de cómo aplicar la prueba. El número de centros que eligieron hacer la prueba en papel fue bastante superior a los que eligieron hacerlo online. Esta limitación fue solventada con la técnica de Propensity Score.

La metodología “Puntaje de Propensión”, más conocida como Propensity Score, empleada en este trabajo, nos ha permitido solventar problemas derivados de la no selección aleatoria de la muestra. Pero esta técnica también presenta algunas limitaciones; una de las cuales es la complejidad a la hora de implementar las técnicas matching. En palabras de Cueto y Mato (2005), “*el matching presenta dificultades en su implementación derivadas de la complejidad que supone encontrar pareja para cada observación del grupo de tratamiento, esto es, observaciones en el grupo de control y en*

el grupo de comparación con igual valor para todas las variables X incluidas en el análisis" (p.29).

Una de las primordiales limitaciones que atañen a este estudio se debe a las bases de datos disponibles. Debemos mencionar que los datos con los que hemos trabajado corresponden al año 2011, lo que conlleva que el estudio pueda parecer algo anticuado. Posiblemente, los resultados con datos más recientes ofrezcan una menor brecha digital, consecuencia de la continua exposición de los adolescentes de hoy en día a los recursos tecnológicos (ordenadores, tablets, móviles).

Además, en las bases de datos disponibles únicamente contamos con dos variables de contexto, lo que impide el estudio de posibles efectos relacionados con otras variables de carácter sociodemográficas (como el nivel socioeconómico, el sexo y la edad). Es cierto que reducimos el sesgo con la técnica Propensity Score, pero únicamente de aquellas variables observadas que se han incluido en el estudio. En el caso que nos concierne, tan sólo se han incluido en el modelo como covariables el tipo de centro (público, privado y concertado) y el distrito (Centro, Norte, Sur, Este, Oeste). La elección de estas variables obedece, evidentemente, a los datos disponibles en el estudio, además de a razones teóricas; pero estas covariables no representan al cómputo de variables de carácter sociodemográfico que pueden afectar en el rendimiento de los sujetos y por tanto pueden haber sesgado las muestras generadas.

Por todo ello se desconoce si el buen balance y la conformación de grupos tan homogéneos se mantendrían en caso de haber introducido un gran número de covariables en el modelo, por ese motivo, para futuros estudios sería conveniente asignar aleatoriamente al centro el modo de aplicación de la prueba para evitar tamaños muestrales muy distintos entre grupos, así como la recogida de mayor número de variables de contexto.

Con esta propuesta metodológica planteada en el presente trabajo se aborda una parte pequeña de un conjunto, dado que las covariables incluidas son insuficientes, pero se realiza una pequeña contribución al mismo, incorporando dos covariables que teóricamente son relevantes e influyentes en el rendimiento de los estudiantes. Así, se

ejemplifica y se ofrece un avance teórico y metodológico sólido en el estudio del Funcionamiento Diferencial de las Versiones, donde la utilidad de las técnicas Propensity Score son pertinentes y rigurosas para el correcto estudio comparativo entre versiones.

BIBLIOGRAFÍA

Abad, F. J., Olea, J., Ponsoda, V., y García, C. (2011). Medición en ciencias del comportamiento y de la salud. *Madrid: Editorial Síntesis*.

Abadie, A., & Imbens, G. W. (2012). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics, 29* (1).

Acar, T., & Kelecioglu, H. (2010). Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory and Practice*, *10*(2), 639-649.

Ackerman, T. A. (1992) Didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*(1), 67-91.

Al-Amri, S. (2007). *Computer-based vs. paper-based testing: are they the same?* In: Khandia, F. (ed.). 11th CAA International Computer Assisted Conference: Proceedings of the Conference on 10th & 11th July 2007 at Loughborough University, Loughborough, pp 3-13.

Alderete, A. M. (2006). Fundamentos del análisis de regresión logística en la investigación psicológica. *Evaluar*, *6*, 52-67.

Alderson, J. C. (2000). Technology in testing: The present and the future. *System*, *28*(4), 593-603.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of educational measurement*, *36*(3), 185-198. Recuperado el 02 de Diciembre de 2015 de http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1999.tb00553.x/epdf

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA) (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(3), 387-416.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Applegate, B. (1993). Construction of geometric analogy problems by young children in a computer-based test. *Journal of Educational Computing Research, 9* (1), 61-77.

Arachchi, S. M., Dias, K., & Madanayake, R. S (2014). A Comparison Between Evaluation of Computer Based Testing and Paper Based Testing for Subjects In Computer**.** *International Journal of Software Engineering & Applications (IJSEA), 5*(1), 57-72. Recuperado del 10 de febrero de 2015 de http://airccse.org/journal/ijsea/papers/5114ijsea05.pdf

Arce-Ferrer, A.J. & Guzman, E.M. (2009). Studying the Equivalence of Computer-Delivered and Paper-Based Administrations of the Raven Standard Progressive Matrices Test. *Educational and Psychological Measurement, 69*(5), 855-867.

Arias, E. (2008). *Detección de DIF con Estadísticos Basados en Tablas de Contingencia: El Mantel-Haenszel.* Tesis de Maestría para optar por el título de Magíster de Psicología. Departamento de Psicología, Universidad Nacional de Colombia, Bogotá Colombia. Recuperado el 23 de septiembre de 2014 de http://www.bdigital.unal.edu.co/1641/1/Tesis_Ana_Cristina_Santana.pdf

Armenta, N. G., Pacheco, C. C., y Pineda, E. D. (2008). Factores socioeconómicos que intervienen en el desempeño académico de los estudiantes universitarios de la Facultad de

Ciencias Humanas de la Universidad Autónoma de Baja California. *Revista de investigación en psicología, 11*(1), 153-165.

Arribas, D. (2004). Diferencias entre los test informatizados de primera generación y los test en papel y lápiz: influencia de la velocidad y el nivel de destreza informática. *Acción Psicología, 3,* 91-100.

Association of Test Publishers (ATP, 2002). *Guidelines for computer-based testing.* Washington, DC, ATP.

Attorresi, H., Picón, J., Abal, F., Aguerri, M. y Galibert, M. (2009). Aplicación del modelo LLTM de Fischer al análisis de las fuentes de dificultad de ítemes de razonamiento deductivo. *Interdisciplinaria,* 26 (1), 77-93.

Austin, P. C. (2011). An introduction to Propensity Score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research, 46*(3), 399-424.

Backhoff, E., Ibarra, M. A., y Rosas, M. (1994). Versión computarizada del examen de habilidades y conocimientos básicos. En *Trabajo presentado en el 23º congreso Internacional de Psicología Aplicada. Madrid, España*.

Backhoff, E., Ibarra, M.A., Rosas, M. y Larrazolo, N. (1999). Sistema de evaluación informatizada para el ingreso a la universidad. En J. Olea, V. Ponsoda y G. Prieto (Eds), *Tests Informatizados. Fundamentos y aplicaciones,* (pp.325–342). Madrid: Pirámide.

Backhoff, E., Ramírez, J.L. y Dibut, L. (2005). Desarrollo e implementación del Examen de Ubicación de Matemáticas (EXUMAT). *Revista de la Educación Superior,* XXXIV(136), 19-32.

Bandeira, W. (2002) *Detección del funcionamiento diferencial del ítem (DIF) en test de rendimiento.* Memoria para optar al Título Doctor, Facultad de Educación, Universidad Complutense de Madrid. Recuperado en http://biblioteca.ucm.es/tesis/edu/ucm-t26457.pdf

Bandeira, W. (2003). Descripción de los principales métodos para detectar el funcionamiento diferencial del ítem (DIF) en el área de la evaluación educativa.*Bordón. Revista de pedagogía*, *55*(2), 177-189.

Barbero, I; Vila, E. y Holgado, F.P. (2008). La adaptación de los tests en estudios comparativos interculturales. *Acción Psicológica,* 5, 7-16.

Bartram, D. (2001). *The impact of the Internet on testing: Issues that need to be addressed by a Code of Good Practice*. Internal report for SHL Group plc.

Bartram, D. (2002). *Review model for the description and evaluation of psychological tests.* European Federation of Psychologists Associations (EFPA).

Bartram, D. (2006). Testing on the Internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram y R.K. Hambleton (Eds.) *Computer-Based Testing and the Internet: Issues and Advances*. Chichester, West Susex: Wiley.

Bartram, D., & Hambleton, R.K. (Eds.) (2006). *Computer-based testing and the Internet: Issues and advances.* Chichester: Wiley.

Becker, K. A., & Bergstrom, B. A. (2013). Test Administration Models.*Practical Assessment, Research & Evaluation*, *18* (14).

Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 25-45). Springer Netherlands.

Belloch, C. (2011). *Recursos tecnológicos para la evaluación psicoeducativa. (Las TICs en Logopedia: Audición y Lenguaje).* Valencia, España: Unidad de Tecnología Educativa (UTE), Universidad de Valencia. Recuperado en http://www.uv.es/bellochc/logopedia/NRTLogo3.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*, 1-39.

Bergstrom, B. A. (1992). *Ability Measure Equivalence of Computer Adaptive and Pencil and Paper Tests: A Research Synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, Abril, 1992. http://files.eric.ed.gov/fulltext/ED377228.pdf

Bergstrom, B., & Lunz, M. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Earlbaum.

Binet. A. y Simon, T. (1973). *The development of intelligence in children*. New York: Arno Press.

Brand, N. y Houx, P.J. (1992). MINDS: Toward a computerized Test Battery for Health Psychological and Neuropsychological Assessment. *BehavioralResearch Methods, Instrumentation and Computers, 24*, 385-389.

Breithaupt, K.J., Mills, C.N., & Melican, G.J. (2006). Facing the opportunities of the future. In D. Bartram y R.K. Hambleton (Eds.), *Computer based testing and the Internet* (pp. 219-251). Chichester: John Wiley and Sons.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS RR-01-23). Princeton, NJ: ETS.

British Psychological Society Psychological Testing Centre (2002). *Guidelines for the Development and Use of Computer-based Assessments.* Leicester: British Psychological Society.

British Standards' Institute (2001). *A code of practice for the use of information technology for the delivery of assessments.* London: British Standards' Institute

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Differential Item Functioning. In *Rasch Analysis in the Human Sciences* (pp. 273-297). Springer Netherlands.

Boo, J., & Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences 1.*Psychological reports*, *111*(2), 443-460.

Bunderson, C.V., Inouye, D.K. & Olsen, J.B. (1989). The four generations of computerizad educational measurement. En R. L. Linn (Ed.). *Educational measurement*. Londres: Collier Macmillan Publishers.

Cabras, S. y Tena, J. (2013). Estimación del efecto causal del uso de ordenadores en los resultados de los estudiantes en la prueba de PISA 2012 PISA 2012. En INEE (ed.): *PISA 2012: Resolución de problemas. Informe Español. Volumen II: Análisis secundario*, Madrid: Instituto Nacional de Evaluación Educativa.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Newbury Park, C.A.: Sage.

Čandrlić. S, Ašenbrener. M & Holenko. M (2014). Online vs. Paper-Based Testing: A Comparison of Test Results. *37th Internation Convention MIPRO 2014.* Recuperado el 3 de diciembre de 2015 en http://bib.irb.hr/datoteka/701534.CE_12_2553.pdf

Cardona, T., González, J. M., y De Gutiérrez, E. (1973). Relación entre el nivel socio-económico y el test de Habilidades Mentales Primarias en Barranquilla, Colombia. *Revista*

Latinoamericana de Psicología, *5*(3), 293-301. Recuperado el 7 de noviembre de 2014 en http://www.redalyc.org/pdf/805/80550305.pdf

Caro, D. H., McDonald, J. T., & Willms, J. D. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education/Revue canadienne de l'éducation*, *32*(3), 558-590.

Casé, L., Neer, R., Lopetegui, S., Doná, S., Biganzoli, B., & Garzaniti, R. (2014). Matrices Progresivas de Raven: efecto Flynn y actualización de baremos. *Revista de Psicología*, *23*(2). Recuperado el 5 de septiembre de 2015 en http://www.revistapsicologia.uchile.cl/index.php/RDP/article/viewFile/36144/37842

Castro, M. (2009). La evaluación educativa desde la perspectiva del valor añadido. *ESE. Estudios sobre educación, 16,* I47-I66. Recuperado el 8 de julio de 2015 en http://dadun.unav.edu/bitstream/10171/9985/1/La%20evaluaci%C3%B3n%20educativa.pdf

Chahín-Pinzón, N. (2014). Aspectos a tener en cuenta cuando se realiza una adaptación de test entre diferentes culturas. *Psychologia: avances de la disciplina, 8(2),* 109-112.

Chamorro, R. P. (2014). Sistema experto para calificar pruebas de desarrollo en estudiantes de la Universidad Continental. *Apuntes de Ciencia & Sociedad*, 4 *(2).*

Chapelle, C. (2001). *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.

Chen, D. W. (2015). Metacognitive Prompts and the Paper vs. Screen Debate: How Both Factors Influence Reading Behavior.

Chen, D. W., & Catrambone, R. (2015). Paper vs. Screen Effects on Reading Comprehension, Metacognition, and Reader Behavior. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 332-336. SAGE Publications.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2015). *Package 'Lordif'*. Recuperado el 20 de agosto de 2015 en https://cran.r-project.org/web/packages/lordif/lordif.pdf

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of statistical software*, *39*(8), 1.

Choi, S.W. & Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K–12 setting*. Paper presented at annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Clariana, R., & Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602.

Cole, N. S. (1993). History and development of DIF. In P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 25–29). New Jersey: Lawrence Erlbaum Associates, Inc.

Coma, M. R. (2012). Técnicas de evaluación de impacto: Propensity Score matching y aplicaciones prácticas con STATA. *Documentos-Instituto de Estudios Fiscales*, (2), 1-58.

Crane, P. K., Belle, G. V., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in medicine*, *23*(2), 241-256.

Cronbach, L. (1990) *Essentials of Psychological Testing*. Estados Unidos: Harper Collins Publisher.

Cueto, I. B., y Mato, D. F. (2005). Evaluación mediante matching de la formación ocupacional: un estudio de caso para España. *Trabajo presentado al XXX Simposio de Análisis Económico, Murcia.* (Vol. 20200, No. 5).

Cuevas, M. L (2013). *Sesgo cultural en los ítems de las pruebas del examen saber 11° en Colombia*. Tesis para optar al título de Magíster en Psicología, Universidad Nacional de Colombia, Facultad de Ciencias Humanas.

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology,106* (3), 639.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological Issues for Computer-Based Assessment. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 143-230). Dordrecht: Springer.

Davey, T. (2005). Computer-based testing. In B. S. Everitt y D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Hoboken, NJ: Wiley.

DECRETO 22/2007, de 10 de mayo, del Consejo de Gobierno, por el que se establece para la Comunidad de Madrid el currículo de la Educación Primaria. Recuperado el 4 de abril de 2015 en http://www.madrid.org/dat_capital/loe/pdf/curriculo_Primaria_madrid.pdf

DECRETO 23/2007, de 10 de mayo, del Consejo de Gobierno, por el que se establece para la Comunidad de Madrid el currículo de la Educación Secundaria Obligatoria. Recuperado el 4 de abril de 2015 en http://www.madrid.org/dat_capital/loe/pdf/curriculo_Secundaria_madrid.pdf

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932-945. Recuperado el 6 de Agosto de 2014 en http://polmeth.wustl.edu/media/Paper/GenMatch.pdf

Dillon, A., (1994). *Designing usable electronic text: Ergonomic aspects of human information usage.* London: Taylor & Francis.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer *Differential item functioning*, (pp. 35-66). Hillsdale NJ: Erlbaum.

Drasgow, F., Luecht, R.M., & Bennett, R.E. (2006). Technology and testing. En R.L. Brennan (Ed.), *Educational measurement*. Westport, CT: ACE/Praeger.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Psychology Press.

Eells, K.; Davis, A., Havighurst, R. J., Herrick V. E & Tyler, R. W. (1951). *Intelligence and cultural differences*. Chicago: University of Chicago Press.

Eleanor, M; Kecskemety. K.M; Ashley, K. E & Morin, B (2014). *Comparing Student Performance on Computer-Based vs. Paper-Based Tests in a First-Year Engineering Course. 360º of Engineering Education*. 121st ASEE Annual Conference & Exposition. June 15-18, 2014.

Elosua, P. (2006). Funcionamiento diferencial del ítem en la evaluación internacional PISA. Detección y comprensión. *RELIEVE*. *12*(2), 247-259. Recuperado el 23 de enero de 2014 en http://www.redalyc.org/pdf/916/91612204.pdf

Elosua, P., y López-Jáuregui, A. (2007). Aplicación de cuatro procedimientos de detección del funcionamiento diferencial sobre ítems politómicos. *Psicothema*, *19*(2), 329-336.

Escorial, S., & Navas, M. J. (2006). Analysis of the gender variable in the EDTC using differential item functioning techniques. *Psicothema*, *18*(2), 319-325.

Fernández, A. L., Pérez, E., Alderete, A. M., Richaud, M. C., y Fernández Liporace, M. (2011) ¿Construir o Adaptar Tests Psicológicos? Diferentes Respuestas a una Cuestión Controvertida. *Evaluar*, *10*, 60-74.

Fidalgo, A. y Ferreres, D. (2002). Supuestos y consideraciones en los estudios empíricos sobre el funcionamiento diferencial de los ítems. *Psicothema*, *14*(2), 491-496.

Ferrera, J. M. C., Cebada, E. C., & Chaparro, F. P. (2013). Rendimiento educativo y determinantes según PISA: Una revisión de la literatura en España. *Revista de educación*, (362), 273-297.

Ferrero, S. (2011). *Una comparación más eficiente y precisa entre Bonferroni y Benjamini Hochberg.* Memoria para optar al título de Licenciado en Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Recuperado el 11 de agosto de 2014 en http://www.dc.uba.ar/inv/tesis/licenciatura/2011/ferro

Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (cord). *Psicometría.* (pp. 371-457). Madrid: Editorial Universitas, S.A.

Finger, M. S., & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment*, *11*(1), 58.

Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (2015). *ICILS 2013 Technical Report*. Amsterdam: IEA.

Fulcher, G. (2000). The 'communicative'legacy in language testing. *System*, *28*(4), 483-497.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The Effect of Computer-Based Tests on Racial-Ethnic and Gender Groups. *Journal of Educational Measurement*, *39*(2), 133-147.

Garcés, C. R., Sepúlveda, M. M., y Riquelme, V. C. (2014). Test informatizados y su contribución a la acción evaluativa en educación. *RED. Revista de Educación a Distancia*, (43), 136-152. Recuperado el 20 de agosto de 2014 en http://www.redalyc.org/pdf/547/54732569001.pdf

Gierl, M. J.y Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164–187.

Gil, E. (1992). *El sistema educativo de la Compañía de Jesús. La «Ratio Studiorum»*. Madrid: UPCO.

Gil-Flores, J. (2011). Estatus socioeconómico de las familias y resultados educativos logrados por el alumnado. *Cultura y educación*, *23*(1), 141-154. Recuperado el 18 de julio de 2015 en http://www.mecd.gob.es/dctm/revista-de-educacion/articulosre362/re36211.pdf?documentId=0901e72b816fbab9

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement*, *62*(6), 1053-1067.

Gómez, J. e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología*, *74*, 3-32.

Gómez, J., Hidalgo, M.D. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo, 31*, 75-84.

Gómez, J. y Navas, M.J. (1998). Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica. *Psicothema,* 10 (3), 685-696.

García, J. I. (2009). *Metodología y diseño de estudios para la evaluación de políticas públicas*. Antoni Bosch editor.

Grejda, G. F. (1992). Effects of word processing on sixth graders' holistic writing and revision. *Journal of Educational Research, 85(3),* 144-149.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In Hambleton, R. K.., Merenda, P. F., y Spielberger, C.

(Eds.), *Adapting educational and psychological tests for cross*-cultural assessment (pp. 3-38). London: L.E.A.

Hambleton, R.K., Clauser, B.E., Mazor, M. & Jones, R. (1993). Advances in the detection of differentially functioning test items. *European journal of psychological assessment, 9(1),* 1-18.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross cultural assessments. In Matsumoto, D. & Van de Vijver, F. J. R. (Eds.), *Cross-Cultural Research Methods in Psychology* (pp. 46- 70). New York: Cambridge University Press.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed response science test. *Applied Measurement in Education, 12*, 211-235.

Herrera, A. N. (2005). *Efecto del tamaño de muestra y la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems.* Tesis Doctoral, Facultad de Psicología, Universidad de Barcelona, Barcelona, España.

Hidalgo, M. D., & López-Pina J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Hidalgo, M. D., y López-Pina, J. A. y Sánchez, J. (1997). Error tipo I y potencia de las pruebas chi-cuadrado en el estudio del funcionamiento diferencial de los ítems. Revista de Investigación Educativa, *15*(1), 149 – 170

Hidalgo, M.D., Gómez, J., y Padilla, J.L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem. *Psicothema*, 17, 509-515.

Hidalgo, M. D., Gómez-Benito, J., & Zumbo, B. D. (2014). Binary Logistic Regression Analysis for Detecting Differential Item Functioning Effectiveness of R 2 and Delta Log Odds Ratio Effect Size Measures. *Educational and Psychological Measurement, 74*(6), 927-949.

Higgins, C. A., Gray, G., Symeonidis, P., & Tsintsifas, A. (2005). Automated assessment and experiences of teaching programming. *Journal on Educational Resources in Computing (JERIC)*, *5*(3), 5.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2011). *Package 'MatchIt'*. Recuperado el 4 de febrero de 2015 en https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software 42(8):* 1-28. Recuperado el 4 de febrero de 2015 en www.jstatsoft.org/article/view/v042i08/v42i08.pdf

Holland, P.W., y Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer y H.J. Braun (eds.): *Test validity*, (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (Eds.). (2012). *Differential item functioning*. Routledge.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does It Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2). Recuperado el 15 de agosto de 2015 en http://files.eric.ed.gov/fulltext/EJ843858.pdf

International Test Commission (ITC) (2005). *International Guidelines on Computer-Based and Internet- Delivered Testing*. Recuperado el 7 de febrero de 2014 en http://www.hogrefe.at/fileadmin/redakteure/PDF/international_guidelines_computerbased_internetdelivered.pdf

Ita, M. E., Kecskemety, K. M., Ashley, K. E., & Morin, B. (2014). *Comparing Student Performance on Computer-Based vs. Paper-Based Tests in a First-Year Engineering Course*. Paper presented at 2014 ASEE Annual Conference, Indianapolis, Indiana. Recuperado el 6 de mayo de 2015 en https://peer.asee.org/20188

Ito, K. & Sykes, R.C. (2004). *Comparability of scores from norm-referenced paper-and-pencil and web-based linear tests for grades 4-12*. Paper presented at American Educational Research Association, San Diego, CA. Recuperado el 8 de junio de 2014 en https://www.ctb.com/img/pdfs/raPaperVsWebLinearTests.pdf

Jackel, B. (2014). Item Differential in Computer Based and Paper Based Versions of a High Stakes Tertiary Entrance Test: Diagrams and the Problem of Annotation. In *Diagrammatic Representation and Inference* (pp. 71-77). Springer Berlin Heidelberg.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, *33*(4), 410-422.

Jiménez. E y González. C (2015). *Estudio comparativo entre dos versiones (papel e informatizada) de una prueba de diagnóstico del rendimiento en matemáticas a estudiantes madrileños de educación Secundaria.* Comunicación presentada en el Segundo Congreso Latinoamericano de medición y evaluación educacional (COLMEE). México DF. Recuperado el 15 de octubre de 2015 en http://www.colmee.mx/public/conferences/1/presentaciones/ponenciasdia3/52Estudio.pdf

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist*, *28*(8), 694.

Johnson, M., y Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment*, *4*, 1-35.

Kain, J. F., & Singleton, K. (1996). Equality of educational opportunity revisited. *New England Economic Review*, 87-114. Recuperado el 19 de noviembre de 2015 en https://www.bostonfed.org/economic/neer/neer1996/neer396f.pdf

Kalogeropoulos, N., Tzigounakis, I., Pavlatou, E. A., & Boudouvis, A. G. (2013). Computer-based assessment of student performance in programing courses. *Computer Applications in Engineering Education*, *21*(4), 671-683.

Karay, Y., Schauber, S. K., Stosch, C., & Schüttpelz-Brauns, K. (2015). Computer Versus Paper-Does It Make Any Difference in Test Performance? *Teaching and learning in medicine*, *27*(1), 57-62.

Karkeen, T., Kim, D. I., & Fatica, K. (2010). *Comparability study of online and paper and pencil tests using modified internally and externally matched criteria*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO. Recuperado el 29 de septiembre de 2014 en http://www.measurementinc.com/sites/default/files/Online%20and%20Paper%20and%20Pencil%20Comparability%20Study%20with%20Alternate%20Design.pdf

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills. *Applied Measurement in Education*, *2*, 207-226.

Kim, S.H. y Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, 22, 345-356

Kim, D., y Huynh, H. (2007). Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments. *The Journal of Technology, Learning, and Assessment*, *6*, 1-31.

Kingston, N. M. (2002). *Comparability of scores from computer- and paper-based administrations for students in grades K-8*. 32nd annual Large-Scale Assessment Conference of the Council of Chief State School Officers, Palm Desert, CA.

Kingston, N. M. (2008). Comparability of computer-and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22-37.

Kiplinger, V.L., and R.L. Linn (1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment, 3*(2), 111-133.

Kveton, P., Jelinek, M., Voboril, D., & Klimusova, H. (2007). Computer-based tests: the impact of test design and problem of equivalency. *Computers in Human Behavior, 23*, 32-51.

Kyllonen, P. (1991). Principles for creating a computerized test battery. *Intelligence, 15*, 1-15.

Kyllonen, P. C. (1994). CAM: A theoretical framework for cognitive abilities measurement. In D. K. Detterman (Ed.), *Current topics in human intelligence*. Norwood, N.J.: Ablex.

Kyllonen, P. C. (1996). Is working memory capacity Spearman's g? In I. Dennis y P. Tapsfield (Eds.), *Human abilities: Their nature and measurement*, (pp. 49-75). Hillsdale, NJ: LEA

Laurie, R., Bridglall, B. L., & Arseneault, P. (2015). Investigating the Effect of Computer-Administered Versus Traditional Paper and Pencil Assessments on Student Writing Achievement. *SAGE Open*, *5*(2), 2158244015584616. Recuperado el 1 de marzo de 2016 en: http://sgo.sagepub.com/content/spsgo/5/2/2158244015584616.full.pdf

Leighton, J.P. (2012). Large-scale assessment design and development for the measurement of student cognition. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scale assessment in education: Theory, issues, and practice* (pp. 13-26). Taylor & Francis/Routledge.

Ley Orgánica 2/2006, de 3 de mayo, de Educación. Boletín Oficial del Estado, 4 de mayo de 2006, núm. 106, pp. 17158-17207. Recuperado el 1 de diciembre 2015 en http://www.boe.es/boe/dias/2006/05/04/pdfs/A17158-17207.pdf

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations.

López-Mezquita, M. T. (2005). *La evaluación de la competencia léxica: Test de Vocabulario. Su fiabilidad y validez*. Madrid: Secretaría General Técnica del Ministerio de Educación.

Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2011). A Comparison of Paper and Online Tests Using a Within-Subjects Design and Propensity Score Matching Study. *Multivariate Behavioral Research, 46*, 544-566.

MacCann. R (2005). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology*. *37*, 79-91.

Maestro, M. (2006) La evaluación del sistema educativo, *Revista de Educación, extraordinario*. España: Ministerio de Educación y Ciencia, pp. 315-336.

Magis, D., Beland, S., Raiche, G., & Magis, M. D. (2015). *Package 'difR'*. Recuperado el 15 de agosto de 2014 en ftp://ftp2.uib.no/pub/cran/web/packages/difR/difR.pdf

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 302-321.

Magis, D., & Facon, B. (2013). Item purification does Not always improve DIF detection a counterexample with Angoff's delta plot. *Educational and Psychological Measurement*, *73*(2), 293-311.

Magis, D., & Facon, B. (2014). deltaPlotR: An R Package for Differential Item Functioning Analysis with Angoff's Delta Plot. *Journal of Statistical Software*,*59*(1), 1-19.

Marcenaro, O. (2014). Del lápiz al ordenador: ¿diferentes formas de evaluar las competencias del alumnado? En INEE (ed.): *PISA 2012: Resolución de problemas. Informe Español. Volumen II: Análisis secundario*, Madrid: Instituto Nacional de Evaluación Educativa.

Martínez, U. (1997): *La integración social de los inmigrantes extranjeros en España*. Editorial Trotta: Madrid.

Martínez-Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Mazor, K.M., Hambleton, R.K. & Clauser, B.E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22* (4), 357-367.

Mazzeo, J., & Harvey, A. L. (1998): *The equivalence of scores from automated and conventional versions of educational and psychological tests: a review of the literature.* Princeton, NJ: Educational Testing Service.

McKee, L. M., & Levinson, E. M. (1990). A Review of the Computerized Version of the Self-Directed Search. *The Career Development Quarterly*, *38*(4), 325-333.

Mead, A. D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and- Pencil Cognitive-Ability Tests - a Metaanalysis. *Psychological Bulletin, 114*, 449-458.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, 54, 17–24.

MESE. (2010). *Evaluación de Diagnóstico de la Comunidad de Madrid 2009/2010*. Informe no publicable de Grupo de Medida y Evaluación de Sistemas Educativos.

Mîndrilã, D. (2010). Maximum Likelihood (ML) and Diagonally Weighted Least Squares (DWLS) Estimation Procedures: A Comparison of Estimation Bias with Ordinal and Multivariate Non-Normal Data. *International Journal of Digital Society, 1*(1) 60-66.

Millsap, R. & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297-334

Moon, J. L. (2013). Comparability of Online and Paper/Pencil Mathematics Performance Measures. *Open Access Theses and Dissertations from the College of Education and Human Sciences*. Paper 168. Recuperado el 2 de agosto de 2014 en http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1169&context=cehsdiss

Molina, J.G., Sanmartín, J. y Pareja, I. (2000). Los bancos de ítems en el escenario actual de la medición psicológica con test. *Revista de Psicología General y Aplicada. 53*, 127-145.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (2012). Perspectivas actuales y retos futuros de la evaluación psicológica. En C. Zúñiga (Ed.), *Psicología, sociedad y equidad*. Santiago de Chile: Universidad de Chile.

Muñiz. J, Elosua. P y Hambleton. R (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema 20*(25), 151-157.

Muñiz, J., y Fernández-Hermida, J.R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo, 31*, 108-121.

Muñiz, J., y Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests. *Papeles del Psicólogo, 66*, 63-70.

Muñiz, J. y Hambleton, R.K. (1999). Evaluación psicométrica de los test informatizados. En J. Olea, V. Ponsoda, y G. Prieto (Eds.), *Test informatizados. Fundamentos y aplicaciones*, (pp. 23-52). Madrid: Pirámide.

Navas, M. J. (1999). Un siglo utilizando tests. *Revista Electrónica de Metodología Aplicada*, 4(2), 1-11.

Nichols, L. (1996). Pencil and paper versus word processing: A comparative study of creative writing in the elementary school. *Journal of Research on Computing in Education, 29*, 159-166.

Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: are they equivalent?. *Ergonomics*, *51*(9), 1352-1375.

Núñez, R. M., Hidalgo, M. D., & López, J. A. (2000). Comparación de dos procedimientos de purificación del test para la evaluación del FDI con el estadístico de Lord y con las medidas de área de Rajú. *Psicothema*,*12*(Suplemento), 399-402.

OCDE (2009). *Informe PISA-ERA 2009. Informe español Resumen ejecutivo*. . Madrid: OCDE, Ministerio de Educación, Cultura y Deporte. Recuperado el 8 de noviembre de 2015, en, http://www.educacion.gob.es/dctm/ministerio/horizontales/prensa/notas/2011/20110627-resumen-ejecutivo-informe-espanol-pisa-era-2009.pdf?documentId=0901e72b80d241d7

OECD (2013a). *PISA 2015: Draft reading literacy framework*. Recuperado el 8 de noviembre de 2015, en http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Reading%20Framework%20.pdf.

OCDE (2013b). *PISA 2012 Results: What Students Know and Can Do*. 4 vols. Paris: OECD. Recuperado el 8 de noviembre de 2015, en, http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf

Olea, J., Abad, F. y Barrada, J. (2010). Test informatizados y otros nuevos tipos de test. *Papeles del Psicólogo. 31*, 94-107.

Olea, J. y Hontangas, P. (1999). Test informatizados de primera generación. En J. Olea, V. Ponsoda, y G. Prieto (Eds.), *Test informatizados. Fundamentos y aplicaciones*, (pp.23-52). Madrid: Pirámide.

Olea, J., y Ponsoda, V. (2013). Test adaptativos informatizados. Madrid: UNED Ediciones. Versión digital.

Olea, V. Ponsoda, y G. Prieto (Eds.) (1999). *Test informatizados. Fundamentos y aplicaciones.* Madrid: Pirámide.

Olmedo, A. (2007). Reescribiendo las teorías de la reproducción social: influencia de la clase social en las trayectorias educativa y laboral del alumnado granadino de Secundaria y Bachillerato. *Revista de educación*, (343), 199-200. Recuperado el 4 de agosto de 2015 en http://www.revistaeducacion.mec.es/re343/re343_20.pdf

Oltman, P. (1994). *The effect of complexity of mouse manipulation on performance in computerized testing (ETS-RR-94-22)*. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. 395 023). Recuperado el 23 de noviembre de 2014 en http://files.eric.ed.gov/fulltext/ED395023.pdf

O'Neill, H. F., Sugrue, B., & Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*, 135-157.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning (pp. 255-276). In P. W. Holland y H. Wainer (Eds.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.

ORDEN 3319-01/2007, de 20 de junio, del Consejero de Educación, por la que se regulan para la Comunidad de Madrid la implantación y la organización de la Educación Primaria derivada de la Ley Orgánica 2/2006, de 3 de mayo, de Educación (BOCM de 20 de julio).

ORDEN 3320-01/2007, de 20 de junio, del Consejero de Educación, por la que se regulan para la Comunidad de Madrid la implantación y la organización de la Educación Secundaria derivada de la Ley Orgánica 2/2006, de 3 de mayo, de Educación (BOCM de 20 de julio).

Oregon Department of Education (2007). *Comparability of student scores obtained from paper and computer administrations*. Office of Assessment and Information Services Oregon Department of Education. Recuperado el 9 de febrero de 2014 en http://www.ode.state.or.us.

Ovalle R, C. (2015). Sobre la técnica de Puntajes de Propensión (Propensity Score Matching) y sus usos en la investigación en Educación. *Educación y Ciencia*, 4(43), 80-89.

Owston, R. D. (1991). *Effects of Word Processing on Student Writing in a High Computer Access Environment. Technical Report 91-3.* York University Centre for the Study of Computers in Education North York, Ontario.

Paek, P. (2005). Recent trends in comparability studies. *Pearson Educational Measurement.* Recuperado el 16 de septiembre de 2014 en *http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TrendsCompStudies.pdf*

Pearson (2002). *Final report on the comparability of computer-delivered and paper tests for Algebra I, Earth Science and English*. Austin, TX: Author.

Pearson (2003). *Virginia standards of learning web-based assessments comparability study report – Spring 2002 administration: Online & paper tests*. Austin, TX: Author.

PIAAC (2013). *Programa Internacional para la Evaluación de las Competencias de la Población Adulta, Volumen 1. Informe español.* Madrid: OCDE, Ministerio de Educación, Cultura y Deporte. Recuperado el 29 de octubre de 2015 en http://www.mecd.gob.es/dctm/inee/internacional/piaac/piaac2012.pdf?documentId=0901e72b8181d500

Poggio, J., Glasnapp, D.R., Yang, X., & Poggio, A.J. (2005). Comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*, 1-30.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Available from http://www.jtla.org

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research, 32*(2), 153-166.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127-143.

Potenza, M., y Dorans, N. (1995). DIF assessment for politomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Puhan, G., Boughton, K. y Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized Testing. *The Journal of Technology, Learning, and Assessment, 6*, 4-20.

Rajmil, L., Robles, N., Murillo, M., Rodríguez-Arjona, D., Azuara, M., Ballester, A., y Codina, F. (2015). Preferencias en el formato de cuestionarios y en el uso de Internet en escolares. In *Anales de Pediatría*, *83* (1), 26-32.

Recio, P. (2012). *Equivalencia e invariancia de medida entre grupos: análisis factorial confirmatorio vs teoría de respuesta al ítem*. Memoria para optar al grado de doctor en la Universidad Complutense de Madrid.

Renom, J. (1992). *Diseño de Tests*. Barcelona: IDEA I+D.

Renom, J. y Doval, E. (1999). Tests Adaptativos informatizados: Estructura y desarrollo. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones* (pp. 127- 162). Madrid: Pirámide

Revuelta, J., Ximénez, C. y Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, *63*, 791-808.

Roca, E (2009). Las evaluaciones internacionales, en E. Martín y F. Martínez Rizo (coord.), *Avances y desafíos en la evaluación educativa*, OEI, Madrid.

Rodríguez, A., y Martínez, F. (2003). Aplicaciones informáticas de psicometría en investigación educativa. *Comunicar-Revista Iberoamericana de Comunicación y educación,* 21, 163-166.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Rolls, S. & Feltham, R. (1993): Practical and professional issues in computer-based assessment and interpretation. En M. Smith y V. Sutherland (Eds.), *Professional issues in selection and assessment*. Chichester, England: John Wiley and Sons.

Rosenbaum, P. & Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Efects. *Biometrika, 70*, 41-55.

Roussos, L. A., Schnipke, D. L., y Pashley, P. J. (1999). A generalized formula for the Mantel Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*, 293-322.

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M & Barendse, M (2015). *Package 'lavaan'*. Recuperado el 28 de diciembre de 2015 en https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Rowan. B. E (2010). *Comparability of Paper-And-Pencil and Computer-Based Cognitive and Non-Cognitive Measures in a Low-Stakes Testing Environment*. Doctoral dissertation. James Madison University. Advisor(s) J. Christine Harmes and Joshua T. Goodman.

Rubin, D. (2005). Causal Inference Using Potential Outcomes. *J. Amer. Statist. Assoc.* 100 (469): 322–331.

Rumberger, R. W. (2004). What can be done to prevent and assist school dropouts. *Intervention with children and adolescents: An interdisciplinary perspective*, 311-334.

Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Education Policy Analysis Archives. 7*, 1-47.

Russell, M. y Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on test conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5 (3),* 1-20.

Salazar, M. I. (2011). *Aproximación bayesiana a los contrastes de hipótesis múltiples con aplicaciones a los microarrays.* Tesis Doctoral, Universidad Complutense de Madrid, Madrid.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development* Series (NCES 2005–457). Washington, DC: U.S. Government Printing Office, U.S. Department of Education, National Center for Education Statistics.

Santana A.C. (2009), *Efecto de la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems a través del procedimiento de regresión logística*. Memoria para optar al título de Magíster en Psicología, Universidad Nacional de Colombia, Facultad de Ciencias Humanas.

Sawaki, Y. (2001). Comparability of conventional and computerized test of reading in a second language. *Language learning & Technology, 5*, 38-59.

Schade, N., Hernández P. y Elgueta B. (2005). Ensayo de Aplicación práctica, el Test Informatizado de Memoria Memopoc. *Revista de Psicología, 14*, 73-88.

Schenkman, B., Fukuda, T., & Persson, B. (1999). Glare from monitors measured with subjective scales and eye movements. *Displays, 20*(1), 11-21.

Scheuneman, J. D. & Grima, A. (1997). Characteristics of quantitative Word items associated with differential performance for female and black examines. *Applied Measurement in Education, 10*, 299-319.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*, 849–869.

Shealy, R. & Stout, W. (1993). An item response theory model of test bias and differential test functioning. In W.P. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317- 375.

Segall, D. O. (1997). Equating CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 219-226). Washington, DC: American Psychological Association.

Seisdedos, N. (1999). Tests informatizados en el mercado. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones* (pp. 379-391). Madrid: Ediciones Pirámide.

Seo, D. G. (2013). *Score Comparability Study of Online and Paper-Pencil Administrations Using Propensity Score Matching Models*. Paper presented at 2013 Annual Meeting of the National Council on Measurement in Education, San Francisco, California. Recuperado el 15 de septiembre de 2015 en https://www.michigan.gov/documents/mde/Appendix_Z_Mode_Comparability_Study_451873_7.pdf

Seo, D. G., & De Jong, G. (2015). Comparability of Online-and Paper-Based Tests in a Statewide Assessment Program Using Propensity Score Matching. *Journal of Educational Computing Research*, *52*(1), 88-113.

Sepúlveda Acevedo, F., y Valderrama Riquelme, J. (2014). *Efecto anclaje y redes sociales: cómo la presencia, ausencia y cantidad de" Me Gusta" puede afectar la percepción de los consumidores*. Tesis Doctoral, Universidad de Chile.

Sim, G. & Horton, M. (2005). Performance and Attitude of Children in Computer Based Versus Paper Based Testing. In P. Kommers & G. Richards (Eds.), *Proceedings of EdMedia: World Conference on Educational Media and Technology 2005* (pp. 3610-3614). Association for the Advancement of Computing in Education (AACE).

Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. En Matsumoto, D., & Van de Vijver, F. J. R. (Eds.), *Cross-cultural research methods in psychology* (pp. 216-240). New York: Cambridge University Press.

Sireci, S., y Zenisky, A.L. (2006). Innovative items format in computer-based testing: In pursuit of construct representation. En S.M. Downing y T.M. Haladyna (Eds.), *Handbook of test development*. Hillsdale, NJ: LEA.

Slosson, R. L., Nicholson, C. L., & Hibpshman, T. H. (1991). *Slosson Intelligence Test, Revised (SIT-R3).* Austin, TX: Slosson Education Publications.

Smith, B., & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior, 23,* 1481-1498.

Solak, E. (2014). Computer versus Paper-Based Reading: A Case Study in English Language Teaching Context. *Mevlana International Journal of Education, 4*(1).

Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore, MD: Warwick & York.

Sun, S. Y., Shieh, C. J., & Huang, K. P. (2013). A research on comprehension differences between print and screen reading. *South African Journal of Economic and Management Sciences*, *16*(5), 87.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Sykes, R. C., & Ito, K. (1997). The effects of computer administration on scores and item parameter estimates of an IRT-based licensure examination. *Applied Psychological Measurement, 21,* 51-63.

Taylor, J., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer-familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report #61). Princeton, NJ: Educational Testing Service.

Texas Education Agency (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based test.* Recuperado el 7 de noviembre de 2011 en http://ritter.tea.state.tx.us/student.assessment/resources/techdigest/Technical_Reports/2008_literature_review_of_comparability_report.pdf.

Thissen, D. & Steinberg, L. (1997). A response model for multiple-choice ítems (pp. 51-65). In W. J. Van der Linden y R. K. Hambleton (Eds.), *Handboock of Modern Item Responde Theory*. New York: Springer-Verlag Inc.

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational & Behavioral Statistics, 21*, 110-130.

Van den Branden, K., Depauw, V., & Gysen, S. (2002). A computerized task-based test of second language Dutch for vocational training purposes.*Language testing*, *19*(4), 438-452.

Van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology*, 31, 33-51

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). *Adapting*

educational and psychological tests for cross-cultural Assessment (pp. 39-64). London: L.E.A

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology*, 54, 119-135.

Van der Flier, H., Mellenbergh, G. J., Adèr, H. J. & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131-145.

Van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review*, *5*(2), 134-150.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education*, *7*(1), 53-79.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992). *How review options, administration mode and anxiety influence scores on computerized vocabulary tests*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. TM018547).

Wang, S. (2004*). Online or paper: does delivery affect results? Administration mode comparability study for Stanford diagnostic Reading and Mathematics tests.* San Antonio, TX.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K–12 mathematics test. *Educational and Psychological Measurement*. *67*, 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*. *68*, 5-24.

Wang, H., & Shin, C.D. (2009). Computer-Based & Paper-Pencil Test Comparability Studies. *Pearson Education Test, Measurement & Research Services Bulletin*, Issue 9.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. In *annual meeting of the National Council on Measurement in Education, San Francisco, CA*.

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *WAIS-IV: Wechsler adult intelligence scale*. San Antonio, TX: Pearson.

Westlund, E. (2013). *Propensity Score Methods in Practice: A Guide to R*. Technical manual.

Wheadon & Adams (2007). *The comparability of onscreen and paper and pencil test: no further research required?* Paper presented at the International Association for Educational Assessment. Annual Conference. Baku, Azerbaiijan.

Whitmore, M.L. & Schumacker, R.E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement*, *59*, 910-927.

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1987). Test de Figuras Enmascaradas, adaptación española: manual. *Madrid: Ediciones TEA*, 29-36.

Worrell, J., Lou Duffy, M., Brady, M. P., Dukes, C., & Gonzalez-DeHass, A. (2015). Training and Generalization Effects of a Reading Comprehension Learning Strategy on Computer and Paper–Pencil Assessments. *Preventing School Failure: Alternative Education for Children and Youth*, 1-11.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays* (RR-04-18). Princeton, NJ: Educational Testing Service.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. En P.W. Holland y H. Wainer (Eds.), *Differential item functioning*, 337 &– 347). New Jersey, EE. UU.: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1996). A measure of DIF effect size using logistic regression procedures. *National Board of Medical Examiners, Philadelphia, PA.*

ANEXOS

ANEXO 1: *Guía para la aplicación online*

1. Cómo acceder a la aplicación informática para iniciar la realización de las pruebas

1.1. Pasos previos

Los equipos informáticos que se usen para realizar la Evaluación de Diagnóstico deben tener el sistema operativo actualizado (ver instrucciones a continuación) o tener instalado **.NET Framework**, que se puede descargar de la siguiente página web:

http://www.microsoft.com/downloads/details.aspx?FamilyID=0856eacb-4362-4b0d-8edd-aab15c5e04f5&displayLang=es

Actualización de Windows XP

- Seleccione: Inicio→ Todos los programas→ Windows Update (Ver figura 1)

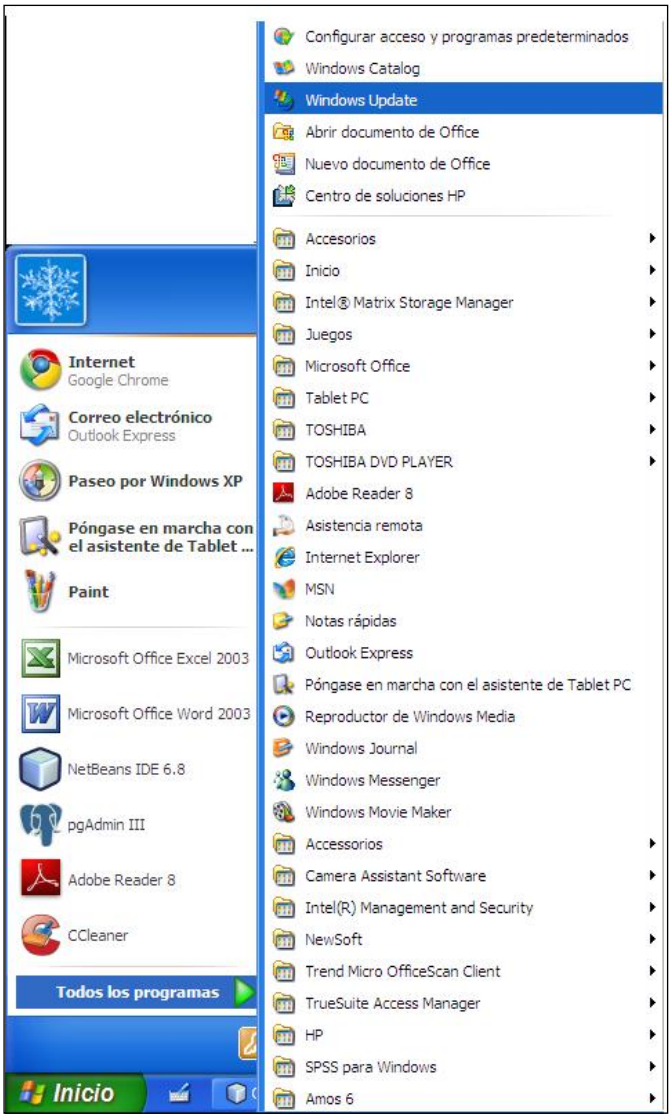


Figura 1

En la página *web* que se abre a continuación, seleccione la opción *Rápida* (figura 2)

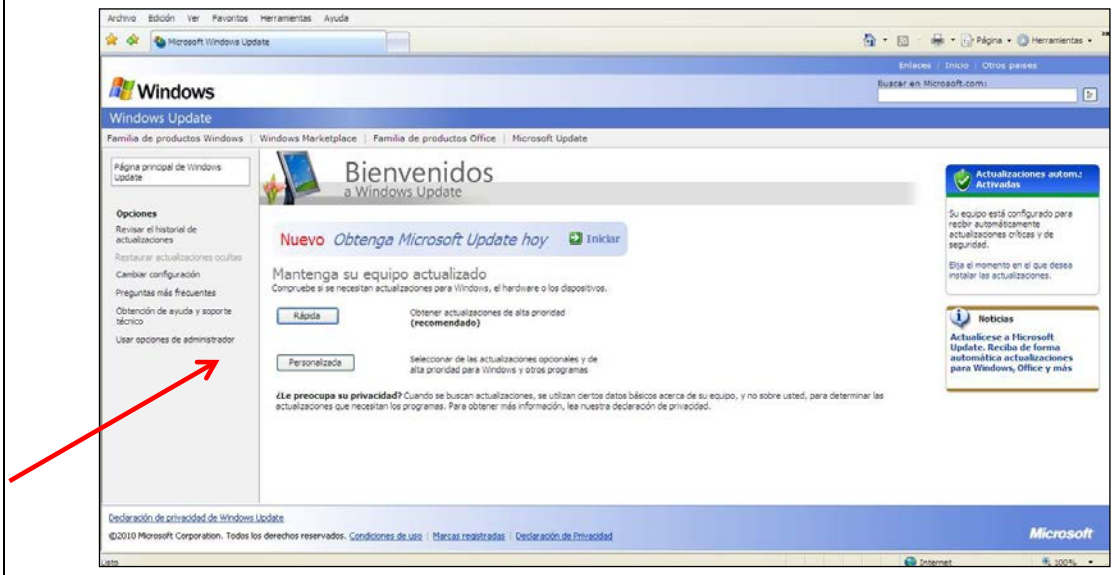


Figura 2

Actualización de Windows Vista y Windows 7

- Seleccione: Inicio→ Todos los programas→ Windows Update
- En el panel izquierdo haga *clic* en *Buscar Actualizaciones* (figura 3)

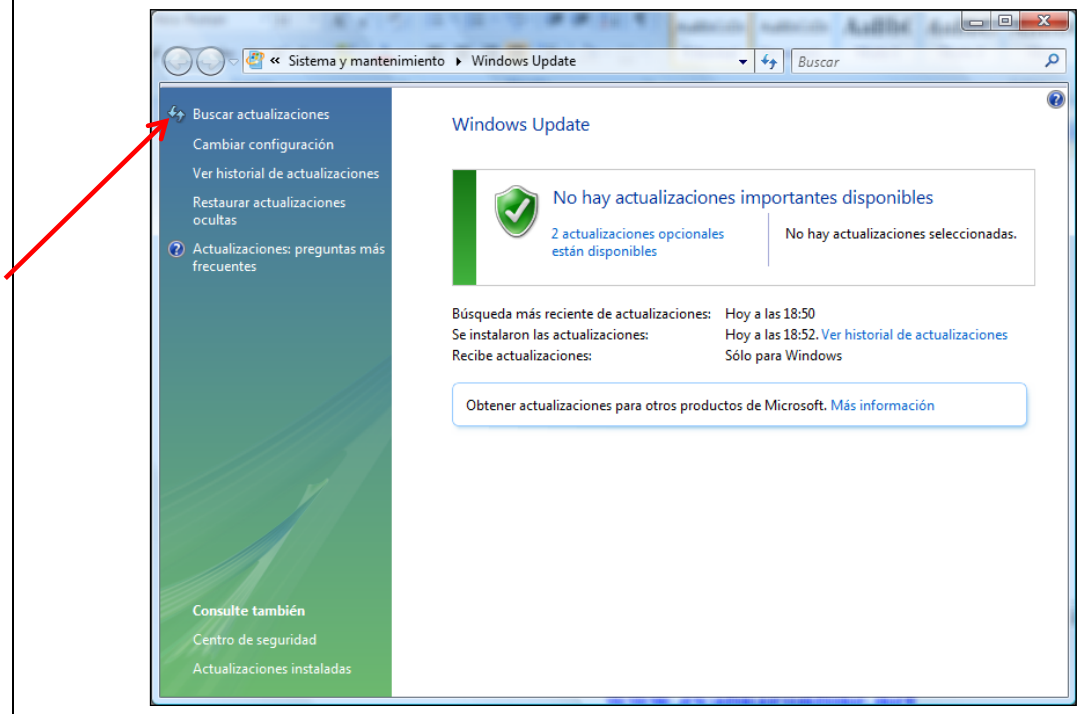


Figura 3

1.2. Instrucciones de acceso a la aplicación

1º La persona responsable de la aplicación de las pruebas debe introducir en cada ordenador la siguiente dirección web:

www.evaluaronline.net

2º En la página web debe seleccionar *Pulse aquí para iniciar la evaluación* (figura 4).

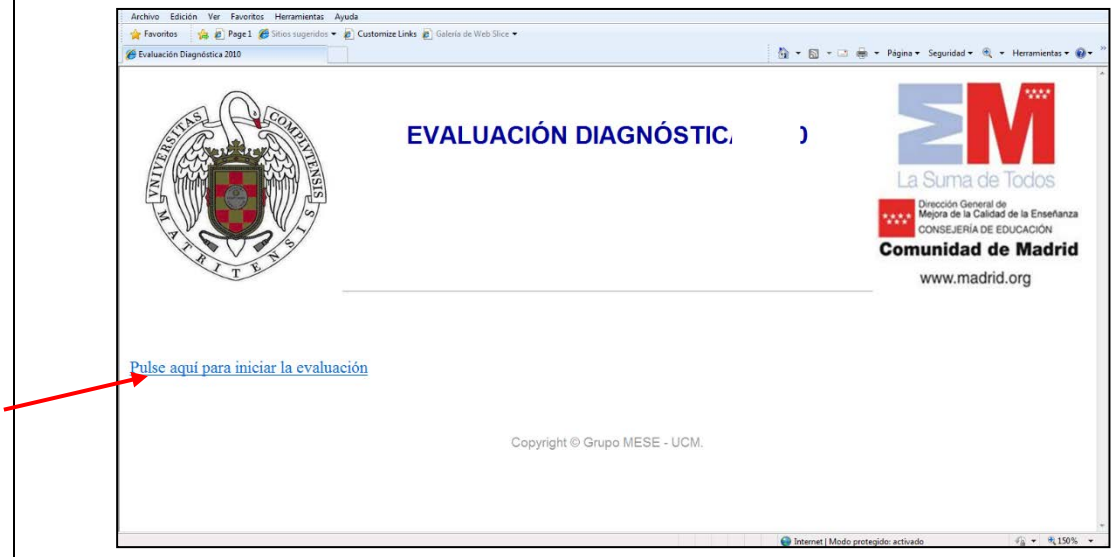


Figura 4

3º En el cuadro de diálogo que se abre inmediatamente a continuación debe pulsar *Ejecutar* (ver figura 5).

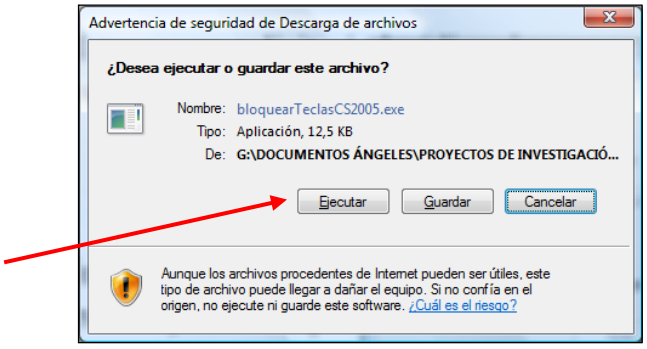


Figura 5

4º Seleccione la opción correspondiente para *iniciar la evaluación* en el cuadro de diálogo que se muestra (figura 6). También está disponible el botón *finalizar la aplicación*, que desbloqueará el teclado cuando se termine de usar el programa.

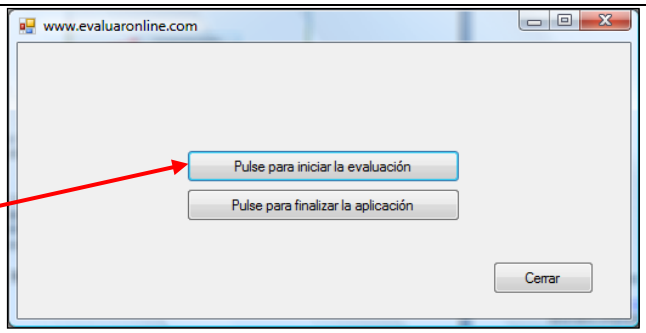


Figura 6

En la figura 7 se reproduce la pantalla inicial de acceso a la aplicación. Tras introducir el código de centro enviado a la dirección, ya se puede comenzar a usar el programa de la Evaluación Diagnóstica 2010-2011 que contiene todas las pruebas consideradas: matemáticas, lengua y comprensión lectora.

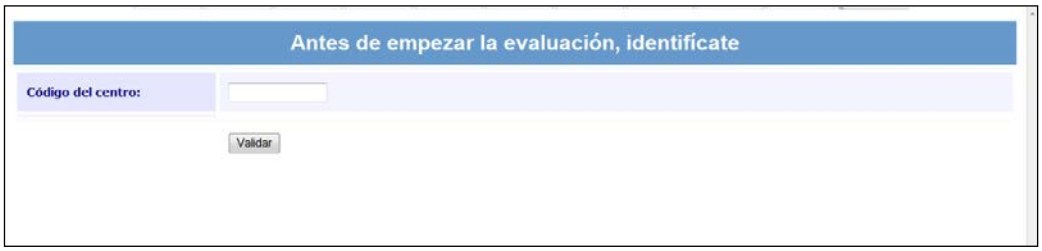


Figura 7

NOTA IMPORTANTE: Por motivos de seguridad es importante que sea el profesor el que instale el programa en cada ordenador, de modo que el alumno inicie su sesión de trabajo sólo cuando la aplicación esté ya en modo de pantalla completa y con el teclado bloqueado, es decir, tal y como se muestra en la figura 7. De este modo los estudiantes no pueden salir de la aplicación, imprimir, copiar, etc. El profesor puede salir de la aplicación en cualquier momento pulsando **ctrl +W**. Cuando *pulse para finalizar la aplicación* el teclado volverá a estar desbloqueado.

2. Realización de las pruebas

1º. El alumno debe seleccionar curso y grupo. Si es su primer acceso y no ha realizado ninguna prueba con anterioridad debe hacer *clic* en el primer botón, en caso contrario, ha de hacer *clic* sobre el segundo (ver figura 8).

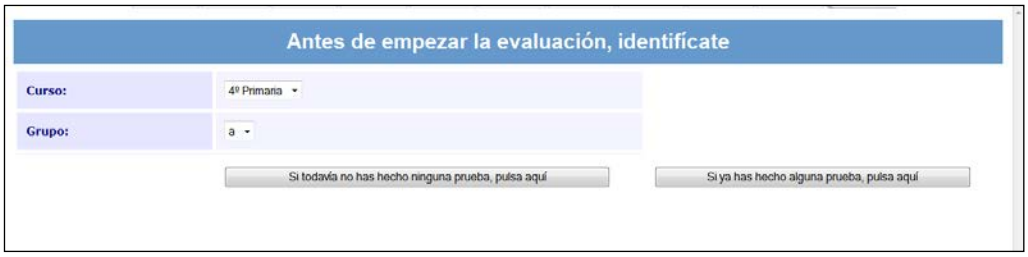


Figura 8

2º En el caso de que se trate del primer acceso del alumno, tras seleccionar el botón *Si todavía no has hecho ninguna prueba, pulsa aquí*, la siguiente pantalla solicitará el nombre y apellidos del alumno. Después de introducirlos, debe pulsar *Entrar (nuevo alumno)*

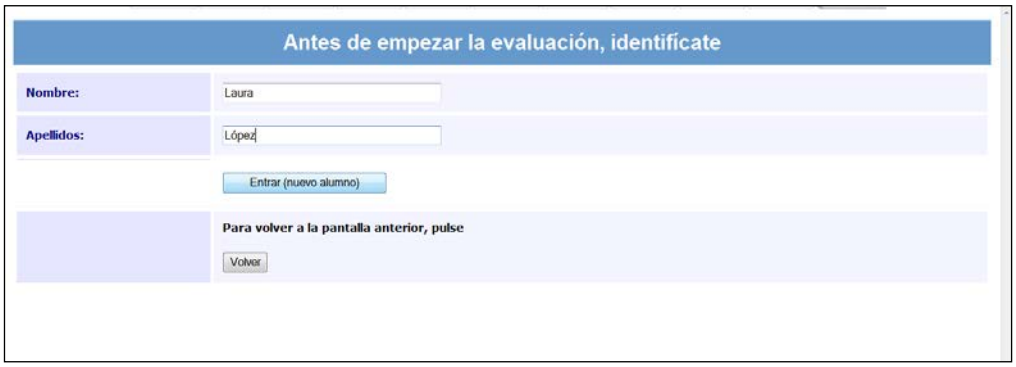


Figura 9

3º A continuación se debe seleccionar la prueba que se vaya a realizar. Puesto que se trata en este caso del primer acceso (el alumno no ha realizado todavía ninguna), aparecen los botones de las tres pruebas disponibles (figura 10).

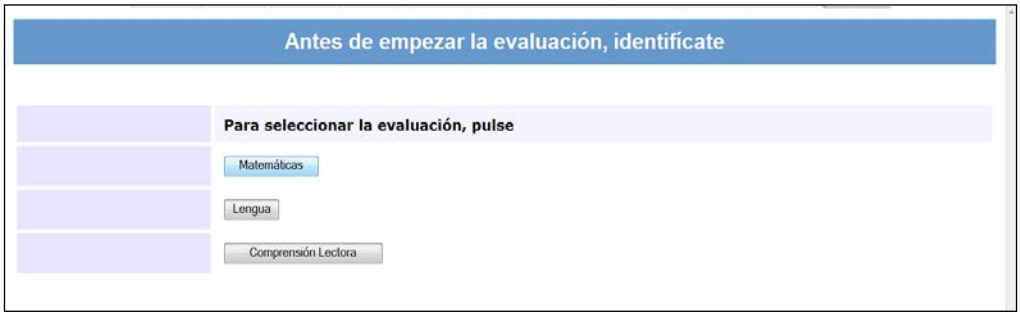


Figura 10

4º Tras seleccionar la prueba deseada, la aplicación solicita que se verifiquen los datos justo antes de *Comenzar la Evaluación*. Si hay algún error, el alumno puede *Volver* atrás para corregirlo (figura 11).

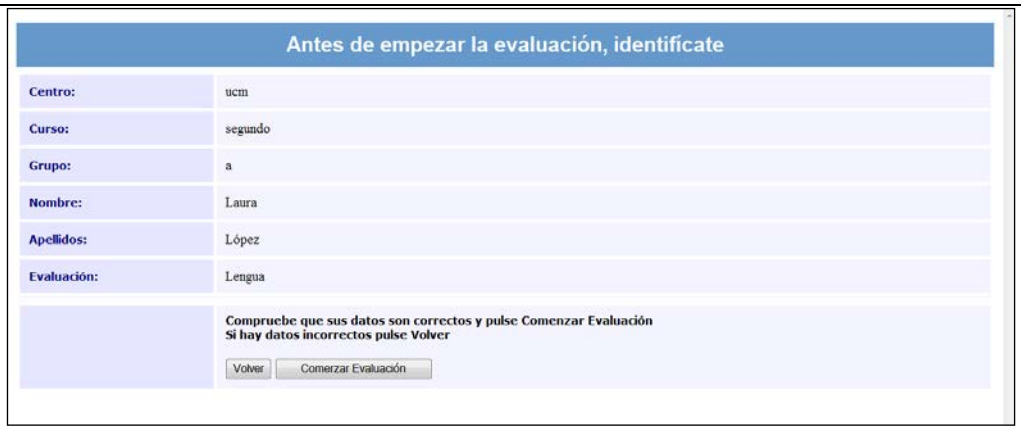


Figura 11

El modo de realización de las pruebas es sencillo y común a todas ellas:

1. Una vez pulsado el botón *Comenzar Evaluación*, se mostrará una primera página de instrucciones. Pulsando al final de las mismas el botón *Siguiente* se accede a la primera pregunta de la prueba.
    a. El alumno puede corregir en todo momento su respuesta a una pregunta pulsando el botón *Limpiar*. Para continuar haciendo la prueba, una vez contestada una pregunta dada, debe seleccionar *Siguiente.*
    b. Salvo en la pantalla de la primera pregunta, el alumno también puede seleccionar *Anterior*, para revisar o cambiar respuestas.
2. Cuando el alumno complete la prueba y haga *clic* en el botón *Siguiente* de la última pregunta, se mostrará una pantalla en la que podrá elegir revisar sus respuestas (*Anterior*) o Finalizar (ver figura *12*).

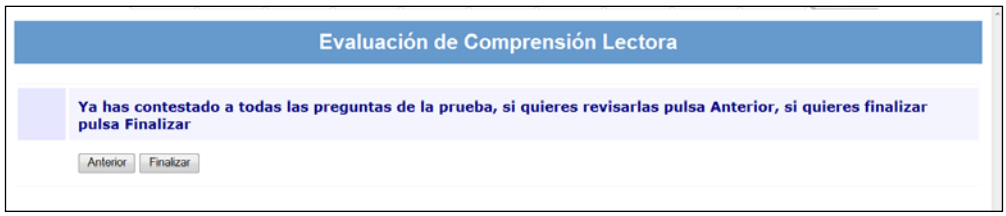


Figura 12

3. Cuando se finaliza una prueba pulsando el botón correspondiente, la aplicación nos vuelve a mostrar la pantalla de introducción de código de centro, lo que nos permite continuar con la evaluación.

NOTA IMPORTANTE: Una vez que un alumno ha pulsado al botón finalizar, y por tanto ha salido de la prueba, ya no podrá acceder a ella de nuevo en ningún momento.

Si el alumno ya ha realizado alguna prueba, al iniciar la aplicación, tras introducir el código de centro, el curso y el grupo, tendrá que seleccionar su

nombre en el listado con el fin de acceder a las pruebas que le queden por realizar (ver figuras 13a y 13b).

Figuras 13a y 13b

La aplicación entonces sólo mostrará las pruebas que el alumno aún no ha realizado. En la figura 14 se muestra, como ejemplo, el caso de un alumno que ya hubiera completado las pruebas de Matemáticas y Comprensión lectora.

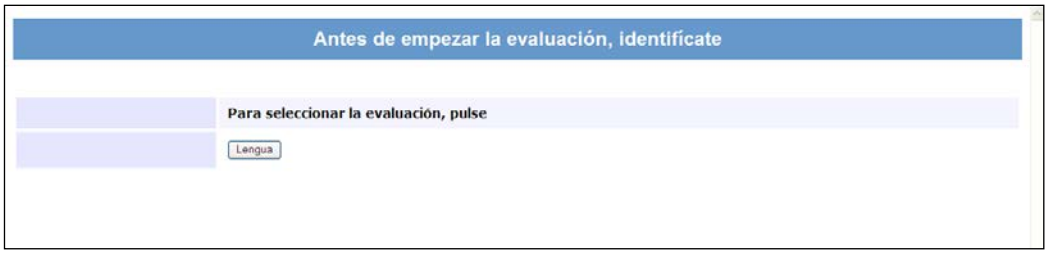


Figura 14

3. Acceso a los resultados obtenidos por el alumnado del centro

En la misma página *web* desde la que accedió a la aplicación, **www.evaluaronline.net**, el profesorado podrá acceder con posterioridad a los resultados obtenidos por los estudiantes de su centro. Para ello tendrá que hacer uso del código de acceso a resultados enviado a la dirección junto con el código de acceso a la aplicación.

ANEXO 2. *Directrices específicas sobre aspectos tecnológicos*

TECHNOLOGY

1. Give due regard to technological issues in Computer-based (CBT) and Internet Testing

a. Give consideration to hardware and software requirements

1. Test Developers

1. Provide a clear description of the minimum hardware and software requirements of the CBT. For Internet testing specify browsers which will support the test.
2. Conduct adequate usability testing of the system requirements using the appropriate delivery platforms to ensure consistency of appearance and delivery.
3. Use appropriate technological features to enhance usability and follow established graphical user interface (GUI) design standards. For example, complex graphics and interactive features may reduce software running speed or increase download time. Items should be designed to fit the test purpose and objectives of assessment, and advanced multimedia features should be used only where justified by validity.
4. Design the system to accommodate likely advances in technology.
5. Design the Internet-delivered testing system to take account of the possibility of fluctuations in demand at different times.
6. Ensure applications of technology advances are tested, documented, and explained to users.
7. Minimise the number of updates and version changes that are issued.
8. Take account of the widely differing connection speeds that apply globally.

2. Test Publishers

1. Verify the documented minimum hardware, software or browser requirements to ensure that they are communicated clearly to the user. Ensure that other technical and operational requirements for the test are explained to the user.
2. Confirm that adequate testing of the system has been completed and documented on the appropriate delivery platforms stated to be suitable.
3. Use only software or hardware features that are essential for measuring the construct and that are likely to be available on systems used by the intended test users and test-takers.
4. Ensure that the test will be as easy as possible to support and maintain in light of likely developments in hardware and software (operating systems etc).
5. Test and document any new features added to the program after publication.

3. Test Users

1. Ensure that you have sufficient understanding of the technical and operational requirements of the test (i.e. hardware and software), as well as the necessary hardware, software and human resources to obtain, use, and maintain the CBT on an on-going basis.
2. Confirm that the system the test-taker is using is documented as being suitable.
3. Ensure there is a good justification for the use of complex software, graphics, and technical IT features in the CBT/Internet test.
4. Monitor supplier for information on future changes to the hardware requirements, test system, or software.
5. Ensure understanding of the implications of changes and their impact on the testing process.

ANEXO 2. *Directrices específicas sobre aspectos tecnológicos (continuación)*

b. Take account of the robustness of the CBT/Internet test

4. Test Developer

1. Test the system to confirm that it is sufficiently robust and capable of dealing with likely system failures and user error.

2. Ensure that the CBT/Internet test is as 'fail-safe' as possible in order to minimise problems arising while the test-taker is responding. Where possible and appropriate:

- treat upper and lower case fonts as equivalent,
- prevent operation of keys or controls that have no function in the test,
- eliminate auto-repeat functions of keys,
- prevent a test-taker from exiting the test by accident,
- provide timely and helpful error feedback,
- follow GUI standards regarding features such as colour, layout, and design, and
- if standardization is not important, allow the user multiple ways to navigate through the system, or allow the user to modify the interface to their liking.

3. When the CBT/Internet test is timed, design the system to respond promptly so that commands have an immediate effect on the screen (e.g., GUI design standards would indicate no more than a 2 second delay onscreen).

4. When the CBT/Internet test is timed, design features so that the time required to move between questions and for the system to record the answer is not part of the timed element (e.g., the test software should deduct these times from the timing of the test or the timing clock should stop during access transitions).

5. For Internet testing, minimise the impact of hang-ups, lost Internet connections and slow downloading (e.g., the system should ensure that no information is lost when the Internet connection is lost).

6. Provide documentation that specifies what to do in the event of routine problems with hardware and/or software.

5. Test Publishers

1. Confirm the robustness of the system has been checked across a range of suitable platforms.

2. Provide sufficient redundancy on all systems throughout the testing site (including incoming and outgoing communications) to allow the site to operate even if one of its components fails.

3. Check the degree to which the test prevents user errors from causing administration problems. Provide users with guidance on what to do in the event that 'bugs' occur during testing (e.g. a test user should be able to report bugs and problems that may be experienced during the testing process).

4. Provide users with contact details (e.g., telephone number, internet address) for technical support.

5. Confirm that the CBT/Internet test responds in a timely manner when taking the test. Where this does not occur, inform test developers and discontinue use of the test until the problem is solved.

6. For Internet testing, put procedures in place to deal fairly with the impact of hang-ups, lost connections and slow downloads. Where download or other technical problems occur, advise the test user/taker of alternatives (e.g., using alternative media or an alternative venue).

7. Document and disseminate relevant technical support to test users. Where appropriate, offer technical support services with trained staff.

ANEXO 2. *Directrices específicas sobre aspectos tecnológicos (continuación)*

6. Test Users

1. Before beginning a test, verify that its robustness has been adequately tested (e.g. documentation provides supporting evidence).
2. Ensure processes are in place to log and resolve problems that may arise during testing.
3. Check availability of the information necessary for contacting the provider of technical support and use technical support services as necessary.
4. Inform test publishers/developers where problems occur with the responsiveness of the computer to the test-taker input.
5. For Internet testing, know the recommended procedures for dealing with hang-ups, lost connections and slow downloads, and advise test-takers accordingly.
6. Provide the test-taker with the technical support specified in the test documentation if any routine problems occur.

c. Consider human factors issues in the presentation of material via computer or the Internet

7. Test Developers

1. Design systems to follow GUI design standards that have been established by groups such as Human Factors International, including but not limited to:
   - ensuring screens have adequate resolution and colour,
   - using consistent screen locations and colour for instructional text and prompts,
   - using consistent screen design, layout and colours,
   - differentiating between test items and test instructions,
   - displaying only relevant information on-screen and ensuring the screen is not overfilled,
   - placing critical information at the start of the text,
   - providing instruction screens with clear fonts and avoiding distracting logos/images,
   - allowing test-takers to review or return to the instruction screen(s) where appropriate, and
   - ensuring representation of status change of display entities (e.g., dimming, highlighting) is consistent in appearance, and logical and meaningful.
2. Display test name, item number, and test prompts or directions at the same location on the screen for each test page.
3. Produce non-alarming, clear and concise error messages that inform how to proceed. Following an error alert, allow the test-taker to correct any errors and continue the test in the most efficient manner possible.

8. Test Publishers

1. Verify that screen design issues have been taken into account in the development of the CBT/Internet test. Where problems are noticed, provide clear and detailed information about the problems to the test developer.
2. Verify that item presentation is consistent throughout the test.
3. Verify that appropriate and informative error messages are presented when necessary.

ANEXO 2. *Directrices específicas sobre aspectos tecnológicos (continuación)*

9. Test Users

1. Be familiar with the screen design requirements of the test and ensure that such features are compatible with the systems being used.
2. Ensure that test-takers are informed of screen design conventions, including where instructional text and prompts are placed, and how instructions can be accessed once testing begins.
3. Be familiar with how items are presented and how the test-taker is required to respond.
4. Verify that error messages are non-alarming and inform how to proceed.

d. Consider reasonable adjustments to the technical features of the test for candidates with disabilities

10. Test Developers

1. Design CBT/Internet tests with hardware/software (e.g., response format) that facilitates the participation of test-takers with disabilities and special needs.
2. Design CBT/Internet tests with hardware and software that can be modified to allow for appropriate test accommodations (e.g., increased font size).

11. Test Publishers

1. Confirm that the hardware/software features of the CBT/Internet test facilitate the participation of test-takers with disabilities and those with special needs (e.g., those who need larger page font).
2. Inform test users about the types of accommodations and modifications that can be made for test-takers with disabilities and those with special needs.
3. Inform test users of the acceptable limits to which tests can be modified or accommodations provided to test-takers.
4. Ensure that test modification and accommodations provided to test users are consistent with legislation regarding individuals with disabilities and special needs.

12. Test Users

1. Check that the hardware/software features facilitate the participation of test-takers with disabilities and those with special needs.
2. Follow best practice as in other modes of testing [see ITC Guidelines on Test Use].
3. Ensure that any necessary test modifications specifically address the test-taker's special needs and are within acceptable limits so as to not adversely affect score validity.
4. Be aware of the impact these modifications may have on the test-taker's score.
5. Consider the use of alternative assessment procedures, rather than modifications to CBT/Internet tests, (e.g., paper and pencil test or alternative structured forms of assessment).

ANEXO 2. *Directrices específicas sobre aspectos tecnológicos (continuación)*

e. Provide help, information, and practice items within the CBT/Internet test

13. Test Developers

1. Provide clear, accurate, and appropriate technical support documentation in both electronic and paper formats. Ensure that such documentation is written at an appropriate level for its target audience.
2. Provide clear instructions on how to load and set up the testing system. For Internet testing, information should be provided on how to log test-takers on and off the system.
3. Provide sufficient and easily available on-screen instructions and help for test-takers. This should include, at a minimum, information about the test (number of items, timing, and types of items) and the testing procedure (how to navigate through the system and how to exit).
4. Where appropriate, develop tutorials or practice tests/items that provide test-takers the opportunity to familiarise themselves with the CBT/Internet test.

14. Test Publishers

1. Provide technical support documentation at a level appropriate for test users. Where appropriate, provide additional customer support services.
2. Disseminate instructions on how to set-up the system to test users. For Internet testing, inform, where appropriate, test users on how to log a test-taker on and off the system.
3. Provide clear and sufficient on-screen instructions.
4. Where appropriate, verify that suitable practice items and tutorials are available. For Internet testing, provide procedures to verify whether a test-taker has accessed practice items and tutorials. Often a test cannot be started until certain practice items have been completed.

15. Test Users

1. Understand the technical support documentation provided with the test and how to access additional technical support when needed.
2. Know how to set up, load and log onto the system.
3. Ensure the test-taker has access to information on the test and the testing process before beginning the test and is able to access on-screen help while completing the test.
4. For Internet testing, provide clear information to the test-taker on how to log-in to and off from the system (e.g., the use of passwords).
5. Provide sufficient opportunity for the test-taker to become familiar with the testing software and the required hardware.
6. Where appropriate, direct test-takers to appropriate Internet testing practice sites.
7. Where appropriate, inform the test-taker about available practice tests. Make it clear that it is the test-taker's responsibility to practice any embedded tutorials and responses to test items (e.g., use of the input device).
8. Where appropriate, collect data on test-taker reactions towards Internet-delivered testing and provide feedback to test developers to help them ensure a more positive experience for test-takers.

Fuente: ITC (2005)

ANEXO 3. *Directrices específicas sobre la calidad*

QUALITY

2. Attend to quality issues in CBT and Internet testing

a. Ensure knowledge, competence and appropriate use of CBT/Internet testing

16. Test Developers

1. Document the constructs that are intended to be measured and investigate whether CBT/Internet mode of delivery is appropriate in terms of content and technical adequacy to access the relevant constructs.
2. Ensure all those involved in test design and development (item writers, psychometricians, software developers etc.) have sufficient knowledge and competence to develop CBT/Internet tests.
3. Remain current on recent advances in CBT/Internet testing, including advances in computer hardware and software technologies and capabilities.
4. Adhere to legal, professional, and ethical mandates and guidelines related to CBT/Internet testing.
5. It is important that during the development of items and tests, the content is protected, through the use of agreements as well as sound security procedures.

17. Test Publishers

1. Ensure that the CBT/Internet test is suitable in terms of content and technical adequacy for its stated purpose and intended test-taker groups.
2. Provide test users with sufficient information about the CBT/Internet test, its modes of operation, and basic computer functions. If appropriate, provide training materials that are specific to CBT/Internet tests and testing.
3. Provide test users with 'best practice' testing policies.
4. Provide test users with clear instructions on how to correctly access and administer Internet tests, including how to log test-takers onto the system.
5. Maintain and regularly update documentation relating to CBT/Internet testing, including pertinent changes in legislation and policy.
6. Adhere to legal, professional, and ethical mandates related to CBT/Internet testing.
7. For Internet testing, document the limitations of the test in terms of the professional context in which it operates:
   - provide a statement indicating the limitations of the relationships between test user and test-taker that can be achieved through this mode (e.g. the Internet is a impersonal medium and a test user may provide only limited advice)
   - provide a statement stating that there are limitations to the conclusions that can be reached just using the Internet test scores.

ANEXO 3. *Directrices específicas sobre la calidad*

b. Consider the psychometric qualities of the CBT/Internet test

19. Test Developers

1. Document and disseminate information on the validity, reliability, and fairness of the CBT/Internet testing process.
2. Ensure that current psychometric standards (test reliability, validity, etc) apply even though the way in which the tests are developed and delivered may differ.
3. Take care that the CBT/Internet test does not require knowledge, skills, or abilities (e.g., computer skills) that are irrelevant to or might impede the test-taker's ability to perform the test.
4. Describe the theoretical and practical applications of algorithms used in test-item selection and/or controlling item or test order (as in adaptive testing).
5. Where test-item content changes, retest and evaluate the changes.

20. Test Publishers

1. Provide appropriate documentation for the psychometric properties of the CBT/Internet test.
2. Ensure that current psychometric standards (test reliability, validity etc.) have been met even though the way in which the tests are developed and delivered may differ.
3. Publish and offer online only those tests that have appropriate psychometric evidence to support their use.
4. When offering assessments online, give advice to test users as to what to look for in order to help them distinguish between tests with and without documented psychometric properties.
5. Verify that the CBT/Internet test does not require knowledge, skills or abilities that are irrelevant to the construct being assessed.
6. Provide documentation that describes the algorithms and measurement models used and present evidence showing that the test has been validated using these algorithms or models.
7. For tests based on models that may be unfamiliar to test users, provide explanations of the relevant concepts for the user.
8. Verify that psychometric model fit has been re-evaluated when changes are made to the test content.

21. Test Users

1. Ensure that documentation of the appropriate psychometric evidence is supplied with the CBT/Internet test.
2. Ensure that current psychometric standards (test reliability, validity etc.) have been met even though the way in which the tests are developed and delivered may differ.
3. Be able to distinguish between tests with and without documented psychometric properties. Those with documented evidence ensure that the evidence is appropriate for the intended use of the test.
4. For Internet testing, use only those websites supported by publishers who offer validated psychometric tests.
5. Check that the CBT/Internet test does not require knowledge, skills or abilities that are irrelevant to the construct being assessed.
6. Where appropriate, review and understand the documentation that describes how the CBT/Internet test uses algorithms for item generation, selection, or test construction, for controlling the order of testing, and the model underlying the development of the test.
7. When necessary, access appropriate training to ensure continuing professional development.
8. Document information provided about changes to test items or parameters and their impact on the test properties.

ANEXO 3. *Directrices específicas sobre la calidad*

c. Where the CBT/Internet test has been developed from a paper and pencil version, ensure that there is evidence of equivalence

22. Test Developers

1. Provide clear documented evidence of the equivalence between the CBT/Internet test and non-computer versions (if the CBT/Internet version is a parallel form). Specifically, to show that the two versions:
   - have comparable reliabilities,
   - correlate with each other at the expected level from the reliability estimates,
   - correlate comparably with other tests and external criteria, and
   - produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores.

2. When designing a CBT/Internet version of a non-computerised test, ensure that:
   - there is equivalent test-taker control (such as the ability to skip or review items) as on the manual version,
   - the method of item presentation ensures that the results from the CBT/Internet test are equivalent to the manual version, and
   - the format for responding is equivalent.

3. For Internet-based tests, studies of test equivalence and norming should be conducted over the Internet with participants completing the test under conditions that represent those that the intended target population will experience (e.g., unproctored or unstandardised testing conditions).

23. Test Publishers

1. Evaluate the documented evidence of the equivalence of the CBT/Internet test, especially if norms from manual versions are to be used by test users to interpret scores on a computerised version of the test.
2. If the developer does not provide evidence of equivalence (e.g., comparable reliabilities, etc.), conduct appropriate equivalence studies.
3. If the developer does not provide evidence relating to the use of the test under conditions that represent those that the intended target population will experience (e.g., unproctored, unstandardised testing), additional studies of test equivalence and norming should be conducted.
4. Verify that the technical features of the CBT/Internet test (e.g., test-taker control and item presentation) allow the results from the CBT/Internet test to be equivalent to the manual version.

24. Test Users

1. Confirm that the evidence regarding the equivalence of the CBT/Internet test to the manual version is sufficient.
2. If norms are based on manual versions of the test, confirm that evidence has been obtained to show equivalence of test means and SDs across versions and for appropriate subpopulations.
3. Verify that the technical features of the CBT/Internet test (e.g., test-taker control and item presentation) allow the results from the CBT/Internet test to be equivalent to the manual version.
4. Only use the test in those modes of administration for which it has been designed (e.g., do not use a test in an unproctored mode when it is specified for use only in proctored modes).

ANEXO 3. *Directrices específicas sobre la calidad (continuación)*

d. Score and analyse CBT/Internet testing results accurately

25. Test Developers

1. Ensure the accuracy of rules/algorithms underlying the scoring of the CBT/Internet test.
2. Provide appropriate documentation of the use and validity of scoring rules.
3. Where reports classify test respondents into categories, such as 'Introverted type' or 'High sales potential', provide information in the test manual that specifies the accuracy of the classification system used to generate computer-based test interpretations (CBTI).
4. Describe the rationale for CBTI statements and how statements are derived from particular scores or score patterns.
5. When test data are hand-entered into a computer, devise procedures to allow for data to be checked for accuracy.

26. Test Publishers

1. Confirm that the accuracy of scoring rules has been adequately evaluated prior to test use.
2. Inform test users about the scoring rules employed within the CBT/Internet test (e.g., use of non-scored items, penalties for guessing).
3. Inform test users how CBTI statements are derived and the validity of that methodology.
4. Stress to test users the importance of carefully checking data input by hand into a computer for scoring.

27. Test Users

1. Review and understand the rules underlying the scoring of the CBT/Internet test.
2. Inform test-takers, when appropriate, about how scores are generated.
3. Know how the statements in the CBTI are derived and be aware of the limitations such methods may have.
4. Ensure the accuracy of test data that are hand-entered into the computer.

e. Interpret results appropriately and provide appropriate feedback

28. Test Developers

1. Illustrate potential limitations of the computer-based test interpretations (CBTI) specific to the current CBT/Internet test.
2. Design and embed individual CBTI report templates for all stakeholders in the testing process.

ANEXO 3. *Directrices específicas sobre la calidad* ***(continuación)***

3. Illustrate how to obtain these various reports and what is contained within each report. In particular consider the:

- media (e.g., text, graphics, etc.),
- complexity of the report,
- report structure,
- purposes of testing,
- degree of modifiability,
- style and tone of report, and
- intended recipients.

4. Provide appropriate guidance on giving feedback, including necessary training requirements for interpreting the CBTI.

29. Test Publishers

1. Inform test users of the potential limitations of interpreting results using CBTI. Specifically:

- statements in a report may be general and not directed towards the specific purpose of the assessment (or specific individuals);
- interpretation is based only on scores of those tests whose data were used as input; therefore, other ancillary data which may be important cannot be taken into account (e.g., scores on other, non-computerised, forms of assessment);
- for open or controlled modes of Internet testing, test-takers may have been tested in non-standardised, unproctored, or variable conditions, whereas score interpretations are based on administration in proctored, standardised conditions;
- some tests are completed in an administration mode that makes it impossible to guarantee the true identity of the test-taker.

2. Assess the suitability of the CBTI provided within the CBT/Internet test system. In particular, take note of

- evidence of the validity and utility of reports,
- the coverage of the reports,
- the consistency of the reports based on similar sets of data,
- the acceptability of the report to intended audiences,
- time, cost and length implications for a test user, and
- freedom from systematic bias.

3. Advise test users on how best to share CBTI with test-takers and other relevant stakeholders.

4. Inform test users of ethical and other accepted practice issues related to providing CBTI feedback to test-takers.

30. Test Users

1. When interpreting the CBTI results, be aware of potential limitations, general and specific, to the reports being used. For example:

- Score interpretations are based on administration in proctored, standardised conditions and the test has been administered under open or controlled modes and there is no evidence provided to support the validity of the report under such conditions.
- Tests are completed in an administration mode that makes it impossible to guarantee the true identity of the test-taker.
- Tests alone, however administered, may not provide a complete assessment of an individual, as other confirmatory or ancillary information is not considered.

ANEXO 3. *Directrices específicas sobre la calidad* ***(continuación)***

2. Select and use the most appropriate CBTI template for the client or intended audience.
3. Ensure that the language and information given in the CBTI fit the needs of the intended stakeholder (e.g., test-taker, organisation, and client).
4. Confirm that there is a sound basis for the CBTI and that its rationale is well-documented.
5. Where possible, edit CBTI reports to include information obtained from other sources to ensure a comprehensive treatment of the test-taker's background, behaviour, ability, aptitude, and personality.
6. Ensure appropriate, relevant, and timely feedback is provided to the test-taker and other relevant stakeholders.
7. Ensure that Internet testing presents test interpretations in a comprehensible and meaningful form.
8. Provide client test interpretations that are appropriate for the context and intended use of the test (e.g., high or low stakes testing, corporate versus individual applications).
9. Take account of ethical issues surrounding the provision of feedback using the Internet (e.g. the difficulty of knowing the effect of providing negative feedback to a test-taker, the lack of knowledge of the emotional state of the test-taker, or the difficulty of providing immediate support to a test-taker when feedback has a negative impact). Where appropriate, feedback should include directions on how to access support and other information.

f. Consider equality of access for all groups

31. Test Developers

1. Document the methods used to enhance psychometric fairness and equality of access.
2. Assess Differential Item Functioning (DIF) and, where DIF might be a problem for one or more groups, identify where this problem occurs and attempt to modify the test to overcome such problems.
3. When developing CBT/Internet tests that may be used internationally, take into account the fact that countries differ in their access to computer technology or the Internet.
4. For tests that are to be used internationally:
   • avoid the use of language, drawings, content, graphics (etc.,) that are country or culture specific.
   • where culture specific tests may be more suitable than culturally-neutral ones, ensure that there is construct equivalence across the different forms.
5. If developing adapted versions of an Internet test for use in specific countries ensure the equivalence of the adapted version and that the adaptation conforms to the ITC Guidelines on Test Adaptation.

ANEXO 3. *Directrices específicas sobre la calidad* ***(continuación)***

32. Test Publishers

1. Where possible, encourage test users to collect biographical data on test-takers in order to monitor the number of people from protected/minority groups who take any CBT/Internet test.
2. Where unequal access to CBT/Internet tests may occur, recommend that test users make alternative forms of assessment available.
3. Inform test users of any evidence regarding DIF for different test-taker groups.
4. When tests are published internationally, provide test users with advice on how to ensure equivalent access to computer technology or the Internet for geographically-diverse groups of test-takers.
5. Where an adapted version of a test is available, provide documentation specifying the equivalence of the adaptation to the original assessment.

33. Test Users

1. To monitor for possible adverse impact, collect data on the number of individuals accessing the CBT/Internet test from protected/minority groups.
• For most countries such groups may be legally defined in terms of one or more of the following: ethnicity, gender, age, disability, religion, and sexual orientation.
2. Where there is evidence of possible inequality of access, offer the use of alternative methods of testing.
3. Where possible, collect data to monitor group differences in test scores.
4. Consider the appropriateness and feasibility of Internet testing if testing in locations with limited access to computer technology or the Internet.
5. If testing internationally, use the country-specific adapted versions of the test, if available.

Fuente: ITC (2005)

ANEXO 4. *Directrices específicas sobre el control*

CONTROL

3. Provide appropriate levels of control over CBT and Internet testing

a. Detail the level of control over the test conditions

34. Test Developers

1. Document the hardware, software, and procedural requirements for administration of a CBT/Internet test.
2. Provide a description of the test-taking conditions required for appropriate CBT/Internet test administration.
3. Design the CBT/Internet test to be compatible with country-specific health and safety, legal, and union regulations and rules (e.g., time on task).

35. Test Publishers

1. Provide sufficient details to test users on hardware, software, and procedural requirements for administering the CBT/Internet test.
2. Describe the test taking conditions candidates should consider when undertaking an Internet-based test.
3. Inform test users of the need to consider health and safety rules during CBT/Internet testing. For example, identify whether an Internet test has the facility for breaks if the testing process is lengthy.

36. Test Users

1. When administering the test, adhere to the standard hardware, software, and procedural requirements specified in the test manual. Before testing, ensure that software and hardware are working properly.
2. When testing at a specific test centre, ensure that the test-taker is comfortable with the workstation and work surface (e.g., the ergonomics are suitable). For example, test-takers should:
   - be encouraged to maintain proper seating posture,
   - be able to easily reach and manipulate all keys and controls,
   - have sufficient leg room, and
   - not be required to sit in one position for too long.
3. When testing via the Internet, provide instructions to test-takers that specify the best methods of taking the test.
4. Ensure that the facilities, conditions, and requirements of the testing conform to national health and safety, and union rules. For example, there may be rules governing the length of time a person should work at a monitor before having a break, or rules as to adequate lighting, heating, and ventilation. When testing over the Internet, inform test-takers of such rules and regulations.

ANEXO 4. *Directrices específicas sobre el control (continuación)*

b. Detail the appropriate control over the supervision of the testing

37. Test Developers

1. Document the level of supervision required for the CBT/Internet test.
• Open mode – No direct human supervision required
• Controlled mode – Although no direct human supervision is required, the test is made available only to known test-takers
• Supervised mode – Test users are required to log on a candidate and confirm that the testing was administered and completed correctly
• Managed mode – A high level of human supervision and control over test-taking conditions is required (as in a dedicated test centre)
2. Provide documentation for the testing scenarios for which the CBT/Internet test has been designed.

38. Test Publishers

1. Document the level of supervision expected for the CBT/Internet test.
2. Specify and restrict the use of specific CBT/Internet tests for particular testing scenarios. For example, psychometric tests for use in post-sift selection testing and/or post-hire assessment normally would not be available in open mode.

39. Test Users

1. Identify the level of supervision required to administer the CBT/Internet test.
2. Use the CBT/Internet test only in the appropriate testing scenarios for which it was designed.

d. Give due consideration to controlling prior practice and item exposure

40. Test Developers

1. For high-stakes Internet-based tests, use software that tries to equate item exposure rates for items drawn from item banks.
2. Limit pilot testing of items on live tests, to minimize unnecessary exposure.
3. Make sure item banks are sufficiently large to permit making multiple parallel forms secure and to manage item exposure rates in adaptive testing.
4. When parallel forms of a test are created, undertake appropriate psychometric analysis to document their equivalence.
5. Contemplate delivery strategies that deter memorization of test content (e.g. by generation of unique tests for each candidate from item banks; or by use of computer adaptive testing).
6. Control exposure of fixed forms in geographies where cheating is more prevalent by restricted administration to supervised or managed modes.

41. Test Publishers

1. Verify that Internet-based maximum performance tests have appropriate controls to reduce item exposure.
2. Provide test users with sufficient information on and training in how to control item exposure.
3. Where appropriate, provide test-takers with practice without compromising the security of the test items.

ANEXO 4. *Directrices específicas sobre el control (continuación)*

42. Test Users

1. Document for test-takers the equivalence of parallel or multiple forms of a test.
2. Protect the CBT/Internet test from previous item exposure by not coaching test-takers with actual test content.
3. Where appropriate, provide test-takers with practice without compromising the security of the actual test items themselves.

d. Give consideration to control over test-taker's authenticity and cheating

43. Test Developers

1. Design features within the system (e.g., the facility for passwords and username access) that enables test publishers/users to have a level of control over access to various parts of the assessment system.

44. Test Publishers

1. Detail the level of authentication required to access various parts of the assessment system, based on the mode of operation used. Exercise control by requiring test users (in the Supervised and Managed modes) and test-takers (in the Controlled mode) to use a username and password when accessing the test.
2. For moderate or high stakes assessment involving multiple stages, provide information on how test users can reduce the risk of test-taker cheating (e.g., having another person to take the test as a proxy). Where an assessment is carried out in open or controlled mode, checks against cheating can be carried out by requiring the test-taker to undertake a subsequent validation assessment in proctored conditions (i.e. supervised or managed conditions) and a comparison of scores made.
3. Identify the threats to test validity that exist if test control is not maintained properly.
4. Provide advice on the design and implementation of 'honesty (honor) policies' in assessment procedures if one or more stages of the process are to be carried out without direct human supervision.

45. Test Users

1. Ensure test-takers provide the appropriate level of authentication before testing begins. Remind test-takers (in the Controlled mode) of the need to obtain a password and username to access the test. In supervised and managed testing conditions, test-takers should be required to provide authentic, government approved picture identification.
2. For moderate or high stakes testing confirm that procedures are in place to reduce the opportunity for cheating. Technological features may be used where appropriate and feasible (e.g., Closed Circuit Television, CCTV) but it is likely that such testing will require the presence of a test administrator, a follow-up supervised assessment, or a face to face feedback session (e.g., for post-sift assessment in job selection situations).
3. For moderate and high stakes assessment (e.g., job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e. at their convenience in non-secure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores.
   - Procedures should be used to check whether the test-taker's original responses are consistent with the responses from the confirmation test.
   - Test-takers should be informed in advance of these procedures and asked to confirm that they will complete the tests according to instructions given (e.g. not seek assistance, not collude with others etc).

   This agreement may be represented in the form of an explicit honesty policy which the test-taker is required to accept.
4. Provide test-takers with a list of expectations and consequences for fraudulent test taking practices, and require test-takers to accept or sign the agreement form indicating their commitment.

Fuente: ITC (2005)

ANEXO 5. *Directrices específicas sobre la seguridad*

SECURITY

4. Make appropriate provision for security and safeguarding privacy in CBT and Internet testing

a. Take account of the security of test materials

46. Test Developers

1. Design features into the CBT/Internet system that minimise the risk of test items, scoring keys, and interpretation algorithms being illegitimately printed, downloaded, copied, or sent electronically to another computer. For example, software can be developed that controls browser function by disabling access to menu selections (such as copy, paste).
2. Design features into the system (e.g., firewalls) that protects the CBT/Internet test system and associated databases from illegal hacking and computer viruses.

47. Test Publishers

1. Protect sensitive features of the test from illegitimate disclosure. For Internet testing, all important intellectual property (e.g., scoring rules, norms, interpretation algorithms) associated with a test should remain on the host server. Only test items and the outputs from report generators usually should appear on the test user's or test-taker's screens.
2. Where appropriate, develop a policy that limits test material access to qualified and authorised test users and testing centres. For example, when testing over the Internet, test users would need to obtain and use a password before they were able to access test materials or set up an assessment for a test-taker.
3. Passwords should be issued only to users qualified to use the Internet test.
4. Verify and check that the CBT/Internet test has features to protect it from illegal hacking and computer viruses. Confirm for Internet testing that reasonable steps have been taken to prevent servers from being accessed by unauthorised or illegal means.
5. For Internet testing, maintain control over the sensitive features of the test and report copyright violations on the Internet. Monitor the web for illegal versions, old/outdated versions and part versions of the Internet test and take steps (e.g., enforcing copyright law) to eliminate these violations.
6. Take steps to secure protection of test content under existing laws.
7. Take appropriate measures to identify stolen test material on the Internet and to estimate its impact of its distribution on the testing program.
8. Take appropriate measures to control the distribution of stolen test material on the Internet including notification of appropriate legal authorities.
9. Maintain a process for the adjudication of security breach allegations and specify appropriate sanctions.

48. Test Users

1. Know the features that have been developed to ensure the security of test materials, and develop procedures that reduce unauthorised access to such materials.
2. Respect the sensitive nature of test materials and intellectual property rights of test publishers/developers.
3. Protect test materials from being copied, printed, or otherwise reproduced without the prior written permission of the holder of the copyright.
4. Protect passwords and usernames from becoming known to others who are not authorised or qualified to have them.
5. Inform the service provider/publisher of any breach in security.

ANEXO 5. *Directrices específicas sobre la seguridad (continuación)*

b. Consider the security of test-taker's data transferred over the Internet

49. Test Developers

1. When designing an Internet test, build in features that safeguard test-taker data and maintain the security of test material transferred over the Internet.
2. Make use of proxy servers, where appropriate, and embed transactions within secure socket layers.
3. Design data management systems to enable users to access, check, and/or delete data from the server in accordance with local data protection and privacy legislation.
4. Design features that ensure regular and frequent backups of all collected data and that allow for recovery of data when problems emerge.

50. Test Publishers

1. Maintain the security of test-taker data transmitted over the Internet (e.g. by encryption).
2. Ensure that test users and test-takers are informed that the host server has correctly received their data.
3. Inform test users of their rights and obligations in relation to local data protection and privacy legislation.
4. Conduct regular and frequent backups of all collected data and provide test users with a detailed disaster recovery plan should problems emerge.

51. Test Users

1. Prior to test administration, have knowledge of and inform test-takers of the security procedures used to safeguard data transmitted over the internet.
2. Confirm with the service provider that they frequently back up data.
3. Verify that the service provider is able to allow test users and authorised others to discharge their responsibilities as data controllers under local data protection and privacy legislation (e.g. the European Union's Directive on Data Protection).

c. Maintain the confidentiality of test-taker results

52. Test Developers

1. Design features to allow secure storage of CBT/Internet test data on computer, disks or server.
2. Maintain the integrity of CBT/Internet test data by providing technology that does not allow unauthorised altering of information and that can detect unauthorised changes to information.
3. Devise encryption devices and password protection that restrict access to test data.

53. Test Publishers

1. When test data must be stored with publishers, specify the procedures and systems to maintain the confidentiality and security of data.
2. Inform test users of who has access to test data, for what purposes, and how long the data will be stored electronically.
3. Adhere to country-specific data protection laws/regulations governing the storage of personal data.

ANEXO 5. *Directrices específicas sobre la seguridad (continuación)*

4. Restrict access to personal data stored on the host server to those who are qualified and authorised.
5. Protect all sensitive personal material held on computer, disk, or a server with robust (non-trivial) encryption devices or passwords.
6. Confirm the security and confidentiality of the backup data when used to store sensitive personal data.

54. Test Users

1. Know how confidentiality will be maintained when data are stored electronically.
2. Adhere to country-specific data protection laws/regulations governing the collection, use, storage and security of personal data.
3. Protect all material via the use of encryption or passwords when storing sensitive personal data electronically on test centre facilities.
4. Apply the same levels of security and confidentiality to backup data as to the data on the live system when backups are used to store personal data.

Fuente: ITC (2005)

ANEXO 6. *Directrices de los test informatizados*

Adaptación APA (1986), Muñiz y Hambleton (1999) y Lorenzo (2003):

La aplicación

- Los efectos sobre las puntuaciones de los test debidos a la aplicación por ordenador, y que no tienen que ver con los objetivos de la evaluación, deberían eliminarse o ser tenidos en cuenta para la interpretación de las puntuaciones.
- Si se introduce algún cambio en el equipamiento estándar, condiciones o procedimientos, respecto a los que se describen en el manual de test o instrucciones de aplicación, debe demostrarse que no afecta de forma apreciable a las puntuaciones de test. De lo contrario, debería llevarse a cabo una calibración adecuada y documentada.
- El entorno en el que se encuentra situada la terminal para realizar el test debe ser tranquilo, confortable y libre de distracciones.
- Los ítems que se presentan en la pantalla deben ser legibles y estar libres de reflejos luminosos.
- El equipamiento debe comprobarse sistemáticamente y mantenerse en condiciones adecuadas. No se debe aplicar el test en un equipamiento defectuoso. Si el equipamiento falla cuando se está aplicando el test, puede haber que volver a aplicar el test completo o parte de él.
- Hay que supervisar la ejecución del test y procurar una asistencia adecuada al examinado si la necesita. Si es técnicamente posible, hay que avisar inmediatamente al responsable cuando ocurra alguna irregularidad.
- Debe entrenarse de forma adecuada en el uso del ordenador a las personas examinadas, así como establecer procedimientos para eliminar cualquier posible efecto sobre las puntuaciones del test que sea debido a la falta de familiaridad del examinado con el ordenador.
- Deben llevarse a cabo ajustes razonables para aquellas personas que pudiesen tener desventajas para realizar el test informatizado. En los casos en los que la desventaja no pueda ser completamente solventada, las puntuaciones obtenidas deben interpretarse con suma prudencia.

La interpretación:

- Los informes generados por ordenador deben utilizarse únicamente en conjunción con el juicio profesional. El usuario debe juzgar la validez del informe automatizado para cada persona evaluada, basándose en su conocimiento profesional de todo el contexto de evaluación y en el rendimiento y características de dicha persona.

Sujeto evaluado:

- La aplicación informatizada de los test debería proporcionar a los examinados al menos el mismo grado de información y control editorial sobre sus respuestas que tendrían si se aplicase el test de forma tradicional.
- Las personas a las que se aplica el test deberían ser claramente informadas de todos los aspectos de la ejecución que sean importantes para el resultado del test.

- El sistema de test informatizado debe presentar el test y la forma de respuesta de modo que no cause una frustración innecesaria o limite la ejecución de los examinados.
- Los servicios de informatización de los test deben establecer los procedimientos oportunos para asegurar la confidencialidad de la información y la privacidad de los examinados.

Propiedades psicométricas

- El sistema de test informatizado debe estar diseñado de modo que su mantenimiento y verificación resulten sencillos.
- El equipo informático, el procedimiento y las condiciones bajo las cuales se obtuvieron los datos para las normas, la fiabilidad y la validez deben describirse con claridad para permitir la replicación exacta de las condiciones.
- Cuando se interpreten puntuaciones de versiones informatizadas de test convencionales, debe establecerse y documentarse la equivalencia de las puntuaciones de las versiones informatizadas antes de proceder a utilizar las normas o puntos de corte obtenidos a partir de los test convencionales. Las puntuaciones provenientes de aplicaciones convencionales y de test informatizados pueden considerarse equivalentes cuando: a) el rango de las puntuaciones de las personas en ambas formas es muy similar; y b) las medias, variabilidad y forma de las distribuciones de las puntuaciones son aproximadamente las mismas o pueden hacerse similares mediante un reescalamiento de las puntuaciones de la versión informatizada.
- Quienes realizan una versión informatizada de un test convencional deben aportar los datos sobre su validez.
- Los servicios de test deberían advertir a los usuarios de los problemas potenciales que pueden existir cuando las puntuaciones de una versión del test no son equivalentes a las de la versión a partir la cual se han establecido las normas.
- El constructor de un test debería proporcionar estudios comparativos de las versiones convencional e informatizada para establecer la fiabilidad relativa de la aplicación informatizada.
- No se puede asumir sin más la precisión de la puntuación y de la interpretación informatizada. Los proveedores de servicios de test informatizados deben comprobar y controlar la calidad del software y del hardware, incluyendo la puntuación, los algoritmos y el resto de los procedimientos descritos en el manual.

Validez de las interpretaciones informatizadas

- Los servicios de test informatizados deben proporcionar un manual en el que se muestren los fundamentos y evidencia que justifican la interpretación de las puntuaciones de la versión informatizada.
- El sistema de clasificación utilizado para elaborar informes de evaluación debe ser suficientemente consistente para el objetivo perseguido. Por ejemplo, en algunos casos es importante que la mayoría de los examinados sean clasificados en los mismos grupos si se les vuelve a pasar el test (asumiendo que la conducta en cuestión no ha cambiado).

- Hay que proporcionar información a los usuarios de los servicios de elaboración de informes automatizados en relación con la consistencia de las clasificaciones, incluyendo, por ejemplo, el número de clasificaciones y el significado que tiene para la interpretación el cambio de una determinada clasificación a las adyacentes.
- Debe facilitarse a los usuarios de los test las puntuaciones originales utilizadas para elaborar interpretaciones informatizadas. La matriz de respuestas originales debe proporcionarse o estar disponible para los usuarios que la soliciten, con la debida consideración para la seguridad del test y la privacidad de los examinados.
- El manual, o en algunos casos el informe evaluativo, debe describir cómo se han derivado de las puntuaciones originales las distintas partes del informe automatizado.
- Los informes evaluativos automatizados deben incluir información sobre la consistencia de las interpretaciones, así como advertir sobre los errores de interpretación más habituales.
- Debe dejarse claro en qué medida las afirmaciones que se hacen en un informe evaluativo automatizado se basan en investigaciones cuantitativas o en una opinión clínica experta.
- Cuando las afirmaciones del informe evaluativo se basan en una opinión clínica experta, hay que proporcionar a los usuarios la información que les permita sopesar la credibilidad de dicha opinión.
- Cuando las predicciones de ciertos resultados o recomendaciones específicas se basan en investigaciones cuantitativas, debe proporcionarse información que muestre las relaciones empíricas entre la clasificación y la probabilidad de la conducta criterio en el grupo de validación.
- Los servicios de tests informatizados deben garantizar que los informes tanto para los usuarios como para los examinados son comprensibles y establecen claramente los límites dentro de los cuales pueden extraerse conclusiones precisas, teniendo en cuenta variables tales como la edad o sexo, las cuales pueden modular las interpretaciones.

Revisión

- Debe facilitarse información adecuada a los profesionales cualificados, implicados en la revisión especializada de los servicios automatizados, sobre el sistema, así como un acceso razonable a la forma de evaluación de las respuestas. Cuando sea necesario proporcionar secretos comerciales, se establecerá un acuerdo escrito para evitar su difusión.

ANEXO 7. *Objetivos de la enseñanza de la Comprensión Lectora en Primaria y Secundaria*

Primaria	Usar los medios de comunicación social y las tecnologías de la información y la comunicación para obtener, interpretar y valorar informaciones y opiniones diferentes. Utilizar la lectura como fuente de placer y de enriquecimiento personal, y aproximarse a obras relevantes de la tradición literaria para desarrollar hábitos de lectura. Comprender textos literarios de géneros diversos adecuados en cuanto a temática y complejidad e iniciarse en los conocimientos de las convenciones específicas del lenguaje literario.
Secundaria	Hacer de la lectura en sus diversos ámbitos (literario, científico, social…) fuente de placer, de enriquecimiento personal y de conocimiento del mundo y consolidar hábitos lectores. Comprender y producir textos literarios utilizando los conocimientos sobre las convenciones de cada género, los temas y motivos de la tradición literaria y los recursos estilísticos. Aplicar con cierta autonomía, los conocimientos sobre la lengua y las normas del uso lingüístico para comprender textos escritos y para escribir y hablar con adecuación (uso de los diferentes registros lingüísticos) coherencia y corrección.

Fuente: *DECRETO 22/2007. B.O.C.M. Núm. 126 (Primaria)*
DECRETO 23/2007. B.O.C.M. Núm. 126 (Secundaria)

ANEXO 8. *Prueba Comprensión Lectora Primaria*

Las pruebas pueden descargarse de la página de la Conserjería de Educación de la Comunidad de Madrid (p.15-30):

http://www.madrid.org/cs/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=filename%3D1%C2%BA+dia.pdf&blobkey=id&blobtable=MungoBlobs&blobwhere=1272030645943&ssbinary=true

En el siguiente enlace puede consultarse las instrucciones para la prueba censal de Primaria:

http://www.educa.madrid.org/web/cp.alarcon.valdemoro/Web/ColePAA10/DocumentosProfesorado/PRUEBAS%20DIAGNOSTICO/2010-2011/Intruccionescensales%204%20ED%202011.pdf

ANEXO 9. *Prueba Comprensión lectora Secundaria*

Las pruebas pueden descargarse de la página de la Conserjería de Educación de la Comunidad de Madrid (p.33-49):

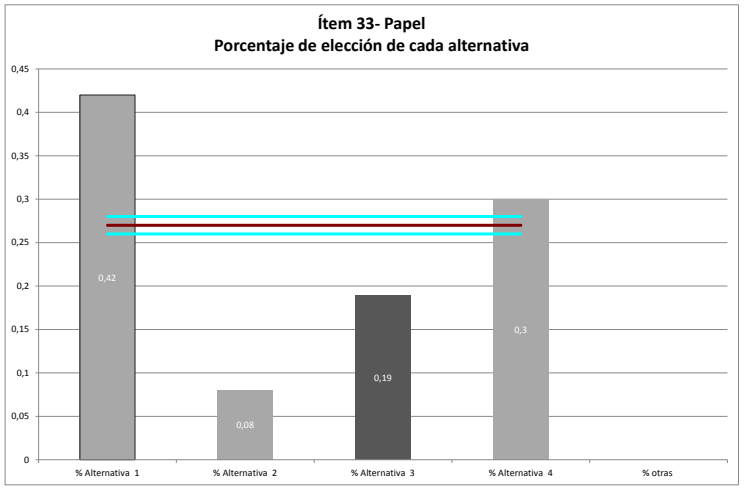
http://www.madrid.org/cs/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=filename%3D2%C2%BA_ESO_ED_2011.pdf&blobkey=id&blobtable=MungoBlobs&blobwhere=1272032113271&ssbinary=true

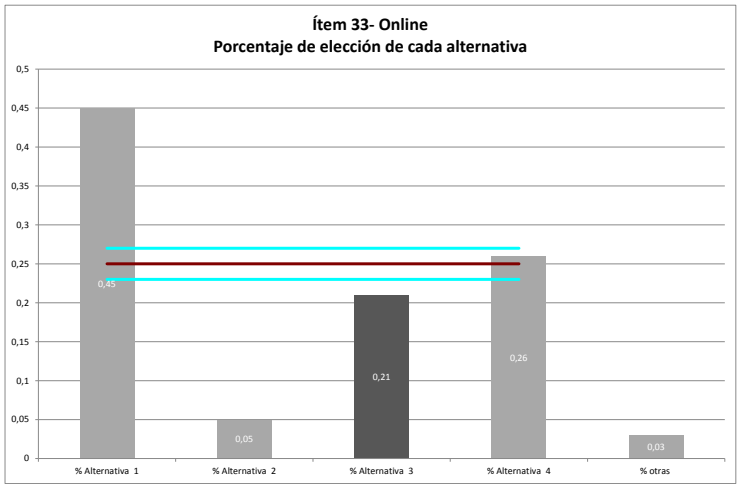
En el siguiente enlace puede consultarse las instrucciones para la prueba censal de Secundaria:

http://www.emprendelo.es/cs/BlobServer?blobkey=id&blobwhere=1272030729168&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=filename%3DInstruccionescensales+2%C2%BA+ESO+ED2011.pdf&blobcol=urldata&blobtable=MungoBlobs

ANEXO 10. *Descripción completa del ítem 33 en la prueba de Comprensión Lectora en Secundaria*

Ítem 33	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial p.corregida	Clave	Media p	Media q	Cov(i,X) (Six)	Cov(i,X) (Six) (Spur)	% Alter. 1	% Alter. 2	% Alter. 3	% Alter. 4	% otras	rbp Alter.1	rbp Alter.2	rbp Alter. 3	rbp Alter. 4	rbp otras	rbp (Spur) Alter. 1	rbp (Spur) Alter. 2	rbp (Spur) Alter. 3	rbp (Spur) Alter. 4	rbp (Spur) otras
Papel	0,19	0,16	0,39	0,07	-0,01	3	24,8	23,9	0,14	-0,02	0,42	0,08	0,19	0,30	0,00	0,01	-0,22	0,07	0,06	0,00	-0,08	-0,27	-0,01	-0,03	0,00
Online	0,21	0,17	0,41	0,12	0,05	3	25,4	23,8	0,27	0,11	0,45	0,05	0,21	0,26	0,03	0,04	-0,13	0,12	0,07	0,00	-0,05	-0,16	0,05	-0,01	0,00





Modelo TRI de 2 Parámetros

Ítem 33	Parámetro a	Parámetro b	p
Papel	0,056	24,085	0,157
Online	-0,266	-5,809	0,015

ANEXO 11. *Tipo de textos para la enseñanza de la Comprensión Lectora en Primaria y Secundaria*

Textos narrativos

- Noticia, crónica, reportaje breve, esquela, cómic, biografía.
- Acta.
- Carta familiar.
- Problema.
- Cuento, novela, romance, fábula, leyenda, biografía, teatro, poema.

Textos descriptivos

- Anuncio por palabras.
- Guía, folleto, ficha.
- Esquema, ficha, c.sinóptico.
- Poema, acotaciones, secuencias de relatos, etopeya, prosopografía, caricatura.

Textos expositivos

- Artículo, agenda, carta al director, cartelera, reseña.
- Anuncio oficial, cartel, ficha, convocatoria.
- Definición, monografía, informe, enciclopedia, artículo científico, libro de texto.
- Poema, secuencia de otros textos.

Textos argumentativos. En el caso de la prueba de 4º de E. Primaria, no aparece dicho tipo de texto por decisiones tomadas en cursos anteriores acerca de la idoneidad de no incluir este tipo de texto para estudiantes de dicha edad.

- Artículo, editorial, anuncio comercial, anuncio institucional, anuncio de propaganda, crítica, entrevista, debate.
- Instancia, panfleto.
- Informe, crítica.

Textos instructivos

- Receta, consultorio.
- Bando, normativa, consigna, guía, folleto de instrucciones, reglas de juego.
- Receta.
- Manual de instrucciones, guía de actividades.
- Acotaciones.

ANEXO 12. *Procesos vinculados con las destrezas de Comprensión Lectora*

	Extraer datos, hechos o informaciones básicas de un texto y relacionarlos con términos y conceptos propios de un determinado campo del conocimiento o la experiencia que pueda identificar o recordar.
Aproximación e identificación	Es un proceso que requiere capacidad para percibir, decodificar, realizar operaciones básicas de inferencia sobre datos explícitos y llevar a cabo una exploración rápida e inicial. De este modo el alumno podrá acceder a una primera comprensión semántica y cognitiva del texto y, activar, en consecuencia el marco general de referencia, la situación o el contexto en que debe interpretarlo y entenderlo.
Organización, síntesis e integración	Organizar la información obtenida reconociendo e identificando partes y relaciones entre ellas que les permitan elaborar esquemas, agrupar la información en bloques funcional o semánticamente afines, reconocer un orden y vincular la información con capos bien acotados y definidos de conocimiento. Sintetizar la información extraída mediante la comparación o el contraste de la información obtenida del texto y de las operaciones anteriores, el examen de las relaciones identificadas y la eliminación de información no relevante. Esta labor de integración permitirá ordenar información dispersa en el texto en secuencias bien conectadas para captar la intención y el sentido global del texto y, en fin, reelaborar la información mediante formas como el resumen, mapas conceptuales y esquemas.
Reflexión y valoración	Se trata de un proceso de reflexión crítica para hacer valoraciones sobre las cualidades del texto en lo referido a la calidad, relevancia, utilizad y eficacia y eficiencia de la información que permite obtener. En el transcurso del proceso es preciso aportar nuevos datos y realizar inferencias complejas en las que entran en juego tanto las informaciones aportadas por el texto como los conocimientos y las experiencias previos del estudiante. Tratándose también de un proceso de valoración aportará datos o ideas que tiendan bien a la convergencia o bien a la divergencia, tanto parciales como totales, con respecto a la información que aporta el texto original y el modo en que lo hace.

ANEXO 12. *Procesos vinculados con las destrezas de Comprensión Lectora (continuación)*

Fuente: Evaluación General de Diagnóstico 2009. Marco de la Evaluación. Ministerio de Educación. Secretaría de Estado de Educación y Formación Profesional. Dirección General de Evaluación y Cooperación Territorial.

	Si la definición del proceso anterior tiene que ver con la capacidad de manejar, reordenar y valorar la información que proporciona un texto vinculando los aspectos formales y de contenido, el proceso de transferencia y aplicación consiste en adaptar, aplicar diseñar, inventar recrear o relacionar la información de modo diferente parara generar nuevos patrones, proponer soluciones alternativas o avanzar incrementando cualitativamente la información que proporciona el texto.
Transferencia y aplicación	El proceso supone, por un lado, que el alumnado ha asimilado la información que le proporciona el texto, por otro, que es capaz de manejar de modo autónomo y, por último, que puede transformarla y aplicarla a situaciones nuevas o diversas. No se trata tan solo de manejar la información que proporciona el texto sino también de utilizarla para aplicarla a situaciones distintas, es decir, de aprehender el sentido del texto y extenderlo a realidades distintas de aquellas que lo motivan. En definitiva, se trata de un proceso de aplicación de conocimiento a partir de una fuente textual concreta y, en ese sentido, la información que proporciona el texto de partida ha de verse globalmente incrementada.

Instituto de Evaluación. Madrid 2009.

ANEXO 13. *Matriz de especificaciones de Comprensión Lectora en Primaria*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
I. NARRATIVOS	APROXIMACIÓN E IDENTIFICACIÓN	- Identificar el argumento - Reconocer los personajes principales - Identificar la secuencia temporal - Distinguir lo que dice el narrador de lo que dicen o piensan los personajes en estilo indirecto - Reconocer las voces de los personajes en los diálogos - Relacionar el título con el argumento - Identificar la moraleja - Entender el titular de noticias o reportajes - Reconocer el papel de la "entradilla" - Identificar el narrador o narradores
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Reconstruir secuencias temporales cuando en el texto hay desorden cronológico - Elaborar titulares aceptables a partir de una noticia sin titular - Reconstruir el orden cronológico de los hechos de la noticia /reportaje - Identificar secuencias descriptivas en noticias, romances, cuentos…
	REFLEXIÓN Y VALORACIÓN	- Valorar si el argumento es realista o fantástico - Juzgar la moraleja - Interpretar el lenguaje figurado: - metáforas - personificaciones - hipérboles
	TRANSFERENCIA Y APLICACIÓN	- Descubrir el tema cuando no está explícito - Realizar inferencias: o Sobre personajes o Sobre el argumento o Sobre espacio y tiempo - Predecir el desarrollo del argumento a partir del título - Deducir el significado de palabras difíciles con ayuda del contexto lingüístico en series (tipo cebras, leones, elefantes, ñus,…) - Deducir el significado aproximado de alguna expresión, con la relectura de la frase completa - Predecir el contenido a partir de titulares - Inferir en contenido en expresiones del tipo: "vuelven a…", "nuevo ataque…". "Ahora…"

ANEXO 13. *Matriz de especificaciones de Comprensión Lectora en Primaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
II. DESCRIPTIVOS	APROXIMACIÓN E IDENTIFICACIÓN	- Identificar párrafo de síntesis o conclusión - Diferenciar rasgos generales ("casa lujosa") de su desarrollo en detalles particulares ("tiene piscina, varios salones, 7 baños…") - Identificar en sustitutos léxicos con carga semántica positiva o negativa - Interpretar elementos icónicos relevantes en fichas, folletos… - Identificar el contenido: qué se describe y se dice de ello.
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Reconocer secuencias descriptivas - Identificar la estructura, el orden seguido en la descripción - Elaborar o completar esquema del contenido - Elaborar o evaluar resúmenes parciales - Desarrollar un cuadro sinóptico de una descripción - Desarrollar un esquema de la descripción
	REFLEXIÓN Y VALORACIÓN	- Interpretar la ironía - Identificar las hipérboles - Identificar las metáforas - Identificar las connotaciones en general - Relacionar los recursos con la actitud del autor - Reconocer el carácter temporal de las descripciones y su reflejo en los rasgos lingüísticos
	TRANSFERENCIA Y APLICACIÓN	- Inferir elementos no citados (época, carácter, rasgos físicos…) - Inferir la intención en descripciones muy incompletas en anuncios

ANEXO 13. *Matriz de especificaciones de Comprensión Lectora en Primaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
III. EXPOSITIVOS	APROXIMACIÓN E IDENTIFICACIÓN	- Identificar la idea literal, la idea principal explícita - Diferencias ideas citadas en estilo directo de las del autor - Identificar la conclusión - Identificar el papel de elementos icónicos relevantes - Identificar el referente de elementos anafóricos relevantes (“por ello, en estos casos..”) - Identificar el referente de sustitutos nominales en sinónimos, hipónimos e hipernónimos - Identificar secuencias descriptivas
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Identificar el tema, reconocer secuencias narrativas - Identificar el tema cuando no está explícito - Identificar ideas Secundarias - Descubrir la estructura - Elaborar o completar el esquema del contenido - Reconocer secuencia temporal si la hay - Reconocer ejemplos - Reconocer el valor de los conectores - Comprender las relaciones causa.-efecto - Identificar la secuencia “problema-solución”, si es el caso
	REFLEXIÓN Y VALORACIÓN	- Valorar si el tema y el título coinciden - Reconocer si se implica o no al receptor - Descubrir posibles incoherencias - Inferir el valor connotativo de ciertas expresiones - Distinguir opiniones de hechos - Valorar si el tema está o no en el título - Reconocer el papel social del emisor, en función de qué escribe: en nombre propio, en función de su cargo, en representación de otros,
	TRANSFERENCIA Y APLICACIÓN	- Anticipar el contenido desde el título o desde el epígrafe - Identificar el contenido desde una lectura parcial - Inferir elementos contextuales que no se mencionan - Identificar la relación emisor-receptor

ANEXO 13. *Matriz de especificaciones de Comprensión Lectora en Primaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
IV. INSTRUCTIVO	APROXIMACIÓN E IDENTIFICACIÓN	- Interpretar órdenes y prohibiciones. - Interpretar las condiciones. - Comprender las excepciones. - Interpretar los indicadores no verbales, como letras, números, guiones. - Interpretar el sentido de los iconos que acompañan a la parte verbal. - Interpretar sangrados, guiones, tipos de letra…, para reconocer la estructura.
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Diferenciar los apartados de que consta el texto: descubrir la estructura. - Identificar secuencias expositivas. - Identificar secuencias descriptivas. - Reconocer la secuencia temporal o sucesión de acciones.
	REFLEXIÓN Y VALORACIÓN	- Distinguir quién es el emisor en acotaciones y diálogos. - Descubrir posibles incoherencias (por ejemplo: añadir un producto X que no se ha señalado en el apartado de “ingredientes”, en una receta). - Diferenciar órdenes de sugerencias o posibilidades.
	TRANSFERENCIA Y APLICACIÓN	- Interpretar correctamente como órdenes expresiones corteses del tipo “por favor”, “se sugiere”, “se ruega”… - Inferir cuestiones de contenido en general.

ANEXO 14. *Matriz de especificaciones de Comprensión Lectora en Secundaria*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
I. NARRATIVOS	APROXIMACIÓN E IDENTIFICACIÓN	- Identificar el significado del léxico - Reconocer el género textual específico - Identificar el argumento - Reconocer los personajes principales - Distinguir lo que dice el narrador de lo que dicen o piensan los personajes en estilo indirecto - Reconocer las voces de los personajes en los diálogos - Relacionar el título con el argumento - Identificar la moraleja - Entender el titular de noticias o reportajes - Reconocer el papel de la "entradilla"
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Reconstruir secuencias temporales cuando en el texto hay desorden cronológico - Reconocer la estructura del contenido - Identificar el narrador o narradores - Reconocer la estructura de la noticia - Elaborar titulares aceptables a partir de una noticia sin titular - Identificar secuencias descriptivas en noticias, romances, cuentos… - Identificar la secuencia temporal
	REFLEXIÓN Y VALORACIÓN	- Valorar si el argumento es realista o fantástico - Juzgar la moraleja - Interpretar el lenguaje figurado: o Metáforas o Personificaciones o Hipérboles - Descubrir la actitud del narrador - Discriminar el tratamiento subjetivo en una noticia
	TRANSFERENCIA Y APLICACIÓN	- Descubrir el tema cuando no está explícito - Realizar inferencias: o Sobre personajes o Sobre el argumento o Sobre espacio y tiempo - Deducir el significado de palabras difíciles con ayuda del contexto lingüístico en series (tipo cebras, leones, elefantes, ñus,…) - Deducir el significado aproximado de alguna expresión, con la relectura de la frase completa - Inferir el contenido en expresiones del tipo: "vuelven a…" , "nuevo ataque…". "Ahora…"

ANEXO 14. *Matriz de especificaciones de Comprensión Lectora en Secundaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
II.DESCRIPTIVOS	APROXIMACIÓN E IDENTIFICACIÓN	- Reconocer el género específico - Identificar el significado del léxico - Identificar párrafo de síntesis o conclusión - Diferencias rasgos generales de su desarrollo en detalles particulares - Identificar el referente en sustituciones anafóricas - Identificar sinónimos, hipónimos e hipernónimos - Identificar sustitutos léxicos con carga semántica positiva o negativa - Interpretar elementos icónicos relevantes. - Identificar el contenido: qué se describe y se dice de ello.
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Identificar la estructura y el orden seguido en la descripción - Elaborar o completar esquema del contenido - Elaborar o evaluar resúmenes parciales - Reformular como anuncio por palabras una oferta desarrollada - Desarrollar un cuadro sinóptico de una descripción - Desarrollar un esquema de la descripción
	REFLEXIÓN Y VALORACIÓN	- Interpretar la ironía - Identificar las hipérboles - Identificar las litotes - Identificar las metáforas - Identificar las connotaciones en general - Relacionar los recursos con la actitud del autor - Reconocer el carácter temporal de las descripciones y su reflejo en los rasgos lingüísticos - Inferir la intención en descripciones muy incompletas en anuncios - Descubrir la actitud el autor. - Descubrir : o Tópicos, o Prejuicios, o Actitudes discriminatorias - Relacionar el léxico utilizado con la intención del autor y su actitud ante lo que describe - Reconocer la intención del autor - Descubrir posibles incoherencias
	TRANSFERENCIA Y APLICACIÓN	- Inferir elementos no citados (época, carácter, rasgos físicos…)

ANEXO 14. *Matriz de especificaciones de Comprensión Lectora en Secundaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
III. EXPOSITIVOS	APROXIMACIÓN E IDENTIFICACIÓN	- Reconocer el género textual específico - Identificar el léxico - Reconocer citas literales - Identificar la idea literal, la idea principal explícita - Diferencias ideas citadas en estilo directo, de las del autor - Identificar la conclusión - Identificar el papel de elementos icónicos relevantes - Identificar el referente de elementos anafóricos relevantes ("por ello, en estos casos…") - Identificar el referente de sustitutos nominales en sinónimos, hipónimos e hipernónimos
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Identificar el tema, reconocer secuencias narrativas en textos expositivos - Identificar el tema cuando no está explícito - Identificar ideas Secundarias - Identificar secuencias descriptivas en el texto expositivo - Descubrir la estructura - Elaborar o completar el esquema del contenido - Reconocer secuencia temporal si la hay - Reconocer ejemplos - Comprender las relaciones causa.-efecto - Reconocer el valor de los conectores - Identificar la secuencia "problema-solución", si es el caso
	REFLEXIÓN Y VALORACIÓN	- Valorar si el tema y el título coinciden - Identificar la actitud del autor - Reconocer si se implica o no al receptor - Descubrir posibles incoherencias - Inferir el valor connotativo de ciertas expresiones - Distinguir opiniones de hechos - Descubrir tópicos, prejuicios o actitudes discriminatorias.
	TRANSFERENCIA Y APLICACIÓN	- Anticipar el contenido desde el título o desde el epígrafe - Identificar desde una lectura parcial - Inferir elementos contextuales que no se mencionan - Reconocer el papel social del emisor, en función de qué escribe: en nombre propio, en función de su cargo, en representación de otros, - Identificar la relación emisor-receptor

ANEXO 14. *Matriz de especificaciones de Comprensión Lectora en Secundaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
IV. ARGUMENTATIVO	APROXIMACIÓN E IDENTIFICACIÓN	- Reconocer el género textual específico. - Identificar el léxico. - Reconocer citas literales. - Identificar las tesis. - Reconocer argumentos. - Reconocer contraargumentos. - Diferenciar ideas citadas en estilo directo, de las del autor. - Interpretar el valor de ciertos conectores (“sin embargo”, “por otra parte”, “por el contrario”…). - Identificar el papel de elementos icónicos relevantes.
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Identificar el tema. - Reconocer la concesión. - Identificar ideas Secundarias. - Descubrir la estructura. - Elaborar o completar el esquema del contenido. - Reconocer ejemplos utilizados como apoyo en la argumentación. - Comprender las relaciones causa- efecto.
	REFLEXIÓN Y VALORACIÓN	- Identificar la actitud del autor. - Reconocer si se implica o no al receptor. - Descubrir posibles incoherencias. - Comprender la ironía. - Identificar las hipérboles. - Identificar los juegos de palabras. - Inferir el valor connotativo de ciertas expresiones. - Distinguir opiniones de hechos. - Diferenciar la información de la sugerencia. - Identificar los reclamos de los anuncios: erotismo, “lo natural”, lo nuevo, la categoría social… - Valorar si el tema está o no está en el título. - Clasificar argumentos (lógicos, de autoridad…). - Evaluar la jerarquía entre los argumentos. - Identificar quién es el receptor ideal y sus rasgos. - Descubrir tópicos, prejuicios o actitudes discriminatorias. - Reconocer el papel social del emisor, en función de qué escribe: en nombre propio, en función de su cargo, en representación de otros…
	TRANSFERENCIA Y APLICACIÓN	- Anticipar el contenido desde el título o desde el epígrafe. - Inferir elementos contextuales que no se mencionan. - Identificar la relación emisor- receptor. - Inferir el contenido de las elipsis en anuncios. - Inferir datos sobre el emisor (ideología, edad, sexo, profesión…)

ANEXO 14. *Matriz de especificaciones de Comprensión Lectora en Secundaria (continuación)*

TIPOS DE TEXTO	PROCESOS	DESCRIPTORES
V. INSTRUCTIVO	APROXIMACIÓN E IDENTIFICACIÓN	- Reconocer el género textual - Identificar el léxico - Reconocer las acotaciones en los textos teatrales. - Interpretar órdenes y prohibiciones. - Reconocer la secuencia temporal o sucesión de acciones. - Interpretar las condiciones. - Comprender las excepciones. - Interpretar los indicadores no verbales, como letras, números, guiones. - Interpretar el sentido de los iconos que acompañan a la parte verbal. - Interpretar sangrados, guiones, tipos de letra…, para reconocer la estructura.
	ORGANIZACIÓN, SÍNTESIS E INTEGRACIÓN	- Diferenciar los apartados de que consta el texto: descubrir la estructura. - Identificar secuencias expositivas. - Identificar secuencias descriptivas.
	REFLEXIÓN Y VALORACIÓN	- Distinguir quién es el emisor en acotaciones y diálogos. - Descubrir posibles incoherencias (por ejemplo: añadir un producto X que no se ha señalado en el apartado de “ingredientes”, en una receta). - Diferenciar órdenes de sugerencias o posibilidades. - Identificar lo más relevante (bien por la tipología, por el lenguaje, por la reiteración…).
	TRANSFERENCIA Y APLICACIÓN	- Interpretar correctamente como órdenes expresiones corteses del tipo “por favor”, “se sugiere”, “se ruega”… - Inferir cuestiones de contenido en general.

ANEXO 15. *Estadístico descriptivo de la prueba de Comprensión Lectora Primaria en papel*

	N	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
					Estadístico	Error típico	Estadístico	Error típico
Ítem1	5486	0,66	0,473	0,223	-0,690	0,033	-1,525	0,066
Ítem2	5486	0,85	0,358	0,128	-1,949	0,033	1,798	0,066
Ítem3	5486	0,90	0,306	0,094	-2,587	0,033	4,696	0,066
Ítem4	5486	0,43	0,495	0,245	0,274	0,033	-1,926	0,066
Ítem5	5486	0,42	0,493	0,243	0,343	0,033	-1,883	0,066
Ítem6	5486	0,69	0,461	0,212	-0,846	0,033	-1,285	0,066
Ítem7	5486	0,86	0,344	0,118	-2,113	0,033	2,467	0,066
Ítem8	5486	0,63	0,484	0,234	-0,523	0,033	-1,727	0,066
Ítem9	5486	0,91	0,285	0,081	-2,881	0,033	6,301	0,066
Ítem10	5486	0,42	0,494	0,244	0,316	0,033	-1,901	0,066
Ítem11	5486	0,36	0,479	0,230	0,598	0,033	-1,644	0,066
Ítem12	5486	0,52	0,500	0,250	-0,077	0,033	-1,995	0,066
Ítem13	5486	0,42	0,494	0,244	0,325	0,033	-1,895	0,066
Ítem14	5486	0,50	0,500	0,250	0,006	0,033	-2,001	0,066
Ítem15	5486	0,83	0,378	0,143	-1,733	0,033	1,004	0,066
Ítem16	5486	0,78	0,413	0,170	-1,368	0,033	-0,130	0,066
Ítem17	5486	0,54	0,498	0,248	-0,174	0,033	-1,970	0,066
Ítem18	5486	0,62	0,487	0,237	-0,474	0,033	-1,776	0,066
Ítem19	5486	0,42	0,493	0,243	0,335	0,033	-1,888	0,066
Ítem20	5486	0,38	0,484	0,235	0,512	0,033	-1,739	0,066
Ítem21	5486	0,52	0,500	0,250	-0,093	0,033	-1,992	0,066
Ítem22	5486	0,80	0,401	0,161	-1,492	0,033	0,227	0,066
Ítem23	5486	0,87	0,341	0,116	-2,143	0,033	2,592	0,066
Ítem24	5486	0,70	0,460	0,211	-0,857	0,033	-1,266	0,066
Ítem25	5486	0,52	0,500	0,250	-0,076	0,033	-1,995	0,066
Ítem26	5486	0,76	0,426	0,182	-1,229	0,033	-0,489	0,066
Ítem27	5486	0,88	0,329	0,108	-2,288	0,033	3,237	0,066
Ítem28	5486	0,78	0,412	0,170	-1,375	0,033	-0,108	0,066
Ítem29	5486	0,75	0,434	0,188	-1,143	0,033	-0,693	0,066
Ítem30	5486	0,73	0,445	0,198	-1,027	0,033	-0,945	0,066
Ítem31	5486	0,87	0,338	0,114	-2,185	0,033	2,774	0,066
Ítem32	5486	0,62	0,486	0,236	-0,491	0,033	-1,759	0,066
Ítem33	5486	0,51	0,500	0,250	-0,020	0,033	-2,000	0,066
Ítem34	5486	0,79	0,404	0,163	-1,459	0,033	0,128	0,066

ANEXO 16. *Estadístico descriptivo de la prueba de Comprensión Lectora Primaria online*

	N	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico	Error típico
Ítem1	1079	0,58	0,493	0,243	-0,333	0,074	-1,893	0,149
Ítem2	1079	0,83	0,376	0,142	-1,755	0,074	1,080	0,149
Ítem3	1079	0,85	0,353	0,124	-2,013	0,074	2,058	0,149
Ítem4	1079	0,31	0,462	0,213	0,834	0,074	-1,306	0,149
Ítem5	1079	0,39	0,488	0,238	0,447	0,074	-1,804	0,149
Ítem6	1079	0,63	0,482	0,232	-0,557	0,074	-1,693	0,149
Ítem7	1079	0,80	0,402	0,161	-1,486	0,074	0,209	0,149
Ítem8	1079	0,56	0,497	0,247	-0,241	0,074	-1,945	0,149
Ítem9	1079	0,87	0,339	0,115	-2,171	0,074	2,716	0,149
Ítem10	1079	0,34	0,475	0,226	0,658	0,074	-1,569	0,149
Ítem11	1079	0,27	0,443	0,196	1,055	0,074	-0,888	0,149
Ítem12	1079	0,46	0,499	0,249	0,147	0,074	-1,982	0,149
Ítem13	1079	0,32	0,465	0,216	0,792	0,074	-1,375	0,149
Ítem14	1079	0,41	0,493	0,243	0,349	0,074	-1,882	0,149
Ítem15	1079	0,82	0,388	0,151	-1,630	0,074	0,657	0,149
Ítem16	1079	0,70	0,459	0,210	-0,873	0,074	-1,241	0,149
Ítem17	1079	0,46	0,499	0,249	0,155	0,074	-1,980	0,149
Ítem18	1079	0,55	0,498	0,248	-0,185	0,074	-1,970	0,149
Ítem19	1079	0,34	0,475	0,226	0,663	0,074	-1,564	0,149
Ítem20	1079	0,27	0,442	0,195	1,066	0,074	-0,865	0,149
Ítem21	1079	0,54	0,499	0,249	-0,151	0,074	-1,981	0,149
Ítem22	1079	0,73	0,444	0,197	-1,045	0,074	-0,910	0,149
Ítem23	1079	0,87	0,340	0,116	-2,159	0,074	2,665	0,149
Ítem24	1079	0,74	0,439	0,193	-1,093	0,074	-0,806	0,149
Ítem25	1079	0,45	0,498	0,248	0,188	0,074	-1,968	0,149
Ítem26	1079	0,70	0,459	0,210	-0,873	0,074	-1,241	0,149
Ítem27	1079	0,83	0,372	0,139	-1,799	0,074	1,238	0,149
Ítem28	1079	0,70	0,458	0,210	-0,882	0,074	-1,224	0,149
Ítem29	1079	0,67	0,472	0,223	-0,702	0,074	-1,510	0,149
Ítem30	1079	0,66	0,474	0,225	-0,671	0,074	-1,552	0,149
Ítem31	1079	0,87	0,333	0,111	-2,244	0,074	3,041	0,149
Ítem32	1079	0,60	0,490	0,240	-0,403	0,074	-1,841	0,149
Ítem33	1079	0,49	0,500	0,250	0,039	0,074	-2,002	0,149
Ítem34	1079	0,83	0,378	0,143	-1,737	0,074	1,020	0,149

ANEXO 17. *Estadístico descriptivo de la prueba de Comprensión Lectora Secundaria papel*

	N	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico	Error típico
Ítem1	14511	0,72	0,451	0,204	-0,954	0,02	-1,091	0,041
Ítem2	14511	0,88	0,325	0,106	-2,336	0,02	3,459	0,041
Ítem3	14511	0,83	0,373	0,139	-1,785	0,02	1,185	0,041
Ítem4	14511	0,8	0,401	0,16	-1,494	0,02	0,233	0,041
Ítem5	14511	0,88	0,327	0,107	-2,309	0,02	3,33	0,041
Ítem6	14511	0,72	0,451	0,203	-0,96	0,02	-1,078	0,041
Ítem7	14511	0,59	0,492	0,242	-0,355	0,02	-1,875	0,041
Ítem8	14511	0,5	0,5	0,25	-0,008	0,02	-2	0,041
Ítem9	14511	0,4	0,491	0,241	0,391	0,02	-1,847	0,041
Ítem10	14511	0,55	0,497	0,247	-0,212	0,02	-1,955	0,041
Ítem11	14511	0,86	0,342	0,117	-2,13	0,02	2,539	0,041
Ítem12	14511	0,79	0,405	0,164	-1,446	0,02	0,092	0,041
Ítem13	14511	0,87	0,331	0,11	-2,259	0,02	3,105	0,041
Ítem14	14511	0,66	0,475	0,226	-0,657	0,02	-1,569	0,041
Ítem15	14511	0,61	0,488	0,238	-0,453	0,02	-1,795	0,041
Ítem16	14511	0,67	0,47	0,221	-0,723	0,02	-1,478	0,041
Ítem17	14511	0,47	0,499	0,249	0,119	0,02	-1,986	0,041
Ítem18	14511	0,81	0,394	0,155	-1,565	0,02	0,45	0,041
Ítem19	14511	0,35	0,476	0,226	0,648	0,02	-1,581	0,041
Ítem20	14511	0,62	0,485	0,235	-0,503	0,02	-1,747	0,041
Ítem21	14511	0,85	0,357	0,128	-1,96	0,02	1,84	0,041
Ítem22	14511	0,66	0,474	0,225	-0,673	0,02	-1,547	0,041
Ítem23	14511	0,88	0,326	0,106	-2,325	0,02	3,408	0,041
Ítem24	14511	0,87	0,331	0,109	-2,267	0,02	3,139	0,041
Ítem25	14511	0,66	0,474	0,224	-0,676	0,02	-1,544	0,041
Ítem26	14511	0,83	0,372	0,138	-1,797	0,02	1,228	0,041
Ítem27	14511	0,82	0,381	0,145	-1,696	0,02	0,877	0,041
Ítem28	14511	0,78	0,416	0,173	-1,332	0,02	-0,225	0,041
Ítem29	14511	0,74	0,437	0,191	-1,11	0,02	-0,769	0,041
Ítem30	14511	0,91	0,29	0,084	-2,811	0,02	5,901	0,041
Ítem31	14511	0,91	0,289	0,083	-2,826	0,02	5,99	0,041
Ítem32	14511	0,72	0,448	0,201	-0,99	0,02	-1,019	0,041
Ítem33	14511	0,19	0,394	0,155	1,561	0,02	0,438	0,041
Ítem34	14511	0,68	0,466	0,217	-0,781	0,02	-1,39	0,041

ANEXO 18. *Estadístico descriptivo de la prueba de Comprensión Lectora Secundaria online*

	N	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico	Error típico
Ítem1	2207	0,75	0,435	0,189	-1,133	0,052	-0,718	0,104
Ítem2	2207	0,89	0,313	0,098	-2,493	0,052	4,218	0,104
Ítem3	2207	0,8	0,402	0,162	-1,481	0,052	0,194	0,104
Ítem4	2207	0,78	0,416	0,173	-1,335	0,052	-0,217	0,104
Ítem5	2207	0,9	0,297	0,088	-2,717	0,052	5,388	0,104
Ítem6	2207	0,72	0,447	0,2	-1,006	0,052	-0,989	0,104
Ítem7	2207	0,58	0,493	0,243	-0,334	0,052	-1,89	0,104
Ítem8	2207	0,52	0,5	0,25	-0,097	0,052	-1,992	0,104
Ítem9	2207	0,35	0,478	0,229	0,612	0,052	-1,627	0,104
Ítem10	2207	0,54	0,499	0,249	-0,157	0,052	-1,977	0,104
Ítem11	2207	0,83	0,376	0,141	-1,755	0,052	1,08	0,104
Ítem12	2207	0,68	0,468	0,219	-0,753	0,052	-1,435	0,104
Ítem13	2207	0,84	0,367	0,135	-1,852	0,052	1,431	0,104
Ítem14	2207	0,56	0,497	0,247	-0,235	0,052	-1,947	0,104
Ítem15	2207	0,51	0,5	0,25	-0,059	0,052	-1,998	0,104
Ítem16	2207	0,71	0,453	0,205	-0,934	0,052	-1,129	0,104
Ítem17	2207	0,46	0,498	0,248	0,179	0,052	-1,97	0,104
Ítem18	2207	0,74	0,438	0,192	-1,105	0,052	-0,779	0,104
Ítem19	2207	0,32	0,467	0,218	0,771	0,052	-1,407	0,104
Ítem20	2207	0,71	0,456	0,207	-0,907	0,052	-1,178	0,104
Ítem21	2207	0,77	0,419	0,175	-1,307	0,052	-0,291	0,104
Ítem22	2207	0,59	0,493	0,243	-0,351	0,052	-1,879	0,104
Ítem23	2207	0,86	0,35	0,122	-2,044	0,052	2,18	0,104
Ítem24	2207	0,91	0,292	0,085	-2,779	0,052	5,73	0,104
Ítem25	2207	0,61	0,487	0,237	-0,468	0,052	-1,782	0,104
Ítem26	2207	0,78	0,415	0,173	-1,342	0,052	-0,2	0,104
Ítem27	2207	0,79	0,404	0,163	-1,461	0,052	0,133	0,104
Ítem28	2207	0,77	0,418	0,175	-1,317	0,052	-0,267	0,104
Ítem29	2207	0,72	0,45	0,202	-0,973	0,052	-1,054	0,104
Ítem30	2207	0,92	0,272	0,074	-3,082	0,052	7,506	0,104
Ítem31	2207	0,89	0,314	0,098	-2,486	0,052	4,182	0,104
Ítem32	2207	0,73	0,445	0,198	-1,021	0,052	-0,958	0,104
Ítem33	2207	0,18	0,387	0,15	1,64	0,052	0,691	0,104
Ítem34	2207	0,65	0,479	0,229	-0,607	0,052	-1,632	0,104

ANEXO 19. *Descripción de los ítems en papel de Primaria atendiendo a la Teoría Clásica de los Test*

Ítem Papel	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial p.corregida	Clave	Media p	Media q	Cov(i,X) (Six)	Cov(i,X) (Six)	% Alter. 1	% Alter. 2	% Alter. 3	% Alter. 4	% otras	rbp Alter.1	rbp Alter.2	rbp Alter. 3	rbp Alter. 4	rbp otras	rbp (Spur) Alter. 1	rbp (Spur) Alter. 2	rbp (Spur) Alter. 3	rbp (Spur) Alter. 4	rbp (Spur) otras
CL1	0,67	0,22	0,47	0,33	0,25	2	23,6	19,8	0,84	0,62	0,25	0,67	0,02	0,07	0,00	-0,27	0,33	-0,14	-0,10	0,00	-0,34	0,25	-0,16	-0,15	0,00
CL2	0,86	0,12	0,35	0,29	0,23	4	22,9	18,5	0,53	0,41	0,02	0,08	0,03	0,86	0,00	-0,17	-0,17	-0,15	0,29	0,00	-0,20	-0,22	-0,18	0,23	0,00
CL3	0,90	0,09	0,30	0,35	0,30	3	22,9	16,7	0,55	0,46	0,02	0,05	0,90	0,03	0,00	-0,20	-0,25	0,35	-0,12	0,00	-0,22	-0,29	0,30	-0,15	0,00
CL4	0,43	0,24	0,49	0,36	0,28	4	24,6	20,6	0,96	0,71	0,27	0,19	0,12	0,43	0,00	-0,21	-0,13	-0,10	0,36	0,00	-0,29	-0,20	-0,16	0,28	0,00
CL5	0,41	0,24	0,49	0,23	0,14	1	23,8	21,3	0,59	0,35	0,41	0,26	0,09	0,24	0,00	0,23	-0,08	-0,20	-0,04	0,00	0,14	-0,16	-0,25	-0,12	0,00
CL6	0,70	0,21	0,46	0,41	0,33	2	23,7	19,0	0,99	0,78	0,11	0,70	0,13	0,05	0,00	-0,24	0,41	-0,20	-0,20	0,00	-0,29	0,33	-0,26	-0,24	0,00
CL7	0,86	0,12	0,34	0,30	0,24	1	23,0	18,2	0,55	0,44	0,86	0,08	0,05	0,00	0,00	0,30	-0,24	-0,15	-0,08	0,00	0,24	-0,29	-0,19	-0,09	0,00
CL8	0,63	0,23	0,48	0,41	0,33	2	24,0	19,4	1,06	0,83	0,18	0,63	0,13	0,06	0,00	-0,23	0,41	-0,27	-0,09	0,00	-0,29	0,33	-0,32	-0,14	0,00
CL9	0,91	0,08	0,28	0,28	0,23	4	22,8	17,5	0,41	0,33	0,04	0,03	0,02	0,91	0,00	-0,19	-0,15	-0,11	0,28	0,00	-0,23	-0,18	-0,13	0,23	0,00
CL10	0,41	0,24	0,49	0,34	0,26	3	24,5	20,8	0,90	0,66	0,34	0,11	0,41	0,14	0,00	-0,14	-0,17	0,34	-0,15	0,00	-0,22	-0,23	0,26	-0,21	0,00
CL11	0,34	0,23	0,47	0,24	0,16	1	24,1	21,4	0,62	0,39	0,34	0,56	0,05	0,05	0,00	0,24	-0,06	-0,21	-0,20	0,00	0,16	-0,15	-0,25	-0,24	0,00
CL12	0,51	0,25	0,50	0,33	0,24	4	24,0	20,5	0,87	0,62	0,18	0,10	0,20	0,51	0,00	-0,24	-0,16	-0,05	0,33	0,00	-0,31	-0,21	-0,13	0,24	0,00
CL13	0,42	0,24	0,49	0,40	0,32	3	24,8	20,5	1,06	0,82	0,21	0,23	0,42	0,15	0,00	-0,14	-0,13	0,40	-0,25	0,00	-0,21	-0,21	0,32	-0,31	0,00
CL14	0,50	0,25	0,50	0,37	0,28	3	24,3	20,4	0,99	0,74	0,34	0,09	0,50	0,08	0,00	-0,19	-0,18	0,37	-0,15	0,00	-0,28	-0,23	0,28	-0,20	0,00
CL15	0,83	0,14	0,38	0,42	0,36	2	23,4	17,4	0,86	0,71	0,05	0,83	0,08	0,04	0,00	-0,23	0,42	-0,27	-0,19	0,00	-0,27	0,36	-0,31	-0,23	0,00
CL16	0,78	0,17	0,41	0,39	0,33	3	23,4	18,4	0,87	0,70	0,12	0,06	0,78	0,04	0,00	-0,20	-0,27	0,39	-0,18	0,00	-0,26	-0,31	0,33	-0,21	0,00
CL17	0,54	0,25	0,50	0,43	0,35	4	24,5	19,8	1,15	0,90	0,16	0,20	0,11	0,54	0,00	-0,19	-0,23	-0,18	0,43	0,00	-0,25	-0,30	-0,24	0,35	0,00
CL18	0,63	0,23	0,48	0,43	0,35	2	24,1	19,3	1,10	0,87	0,17	0,63	0,08	0,12	0,00	-0,30	0,43	-0,16	-0,15	0,00	-0,36	0,35	-0,21	-0,21	0,00
CL19	0,41	0,24	0,49	0,28	0,19	2	24,1	21,1	0,72	0,48	0,17	0,41	0,06	0,36	0,00	-0,19	0,28	-0,21	-0,02	0,00	-0,26	0,19	-0,26	-0,11	0,00
CL20	0,36	0,23	0,48	0,36	0,27	4	24,8	20,9	0,91	0,68	0,23	0,13	0,29	0,36	0,00	-0,10	-0,20	-0,14	0,36	0,00	-0,18	-0,26	-0,22	0,27	0,00
CL21	0,53	0,25	0,50	0,27	0,18	1	23,7	20,8	0,72	0,47	0,53	0,03	0,01	0,42	0,00	0,27	-0,16	-0,17	-0,18	0,00	0,18	-0,19	-0,19	-0,26	0,00
CL22	0,81	0,15	0,39	0,44	0,38	3	23,5	17,5	0,93	0,77	0,04	0,09	0,81	0,06	0,00	-0,24	-0,27	0,44	-0,21	0,00	-0,28	-0,32	0,38	-0,25	0,00
CL23	0,87	0,11	0,33	0,38	0,33	2	23,1	17,0	0,68	0,57	0,05	0,87	0,05	0,03	0,00	-0,24	0,38	-0,20	-0,20	0,00	-0,27	0,33	-0,24	-0,23	0,00
CL24	0,69	0,21	0,46	0,25	0,17	4	23,2	20,3	0,63	0,41	0,05	0,24	0,02	0,69	0,00	-0,26	-0,10	-0,13	0,25	0,00	-0,30	-0,17	-0,16	0,17	0,00
CL25	0,53	0,25	0,50	0,36	0,27	1	24,1	20,3	0,95	0,70	0,53	0,02	0,40	0,05	0,00	0,36	-0,18	-0,23	-0,18	0,00	0,27	-0,20	-0,32	-0,22	0,00
CL26	0,77	0,17	0,42	0,43	0,36	2	23,5	18,1	0,95	0,78	0,05	0,77	0,04	0,14	0,00	-0,23	0,43	-0,23	-0,25	0,00	-0,27	0,36	-0,26	-0,31	0,00
CL27	0,88	0,11	0,33	0,41	0,36	1	23,1	16,4	0,71	0,60	0,88	0,03	0,05	0,03	0,00	0,41	-0,18	-0,24	-0,25	0,00	0,36	-0,22	-0,28	-0,28	0,00
CL28	0,79	0,16	0,41	0,52	0,46	3	23,7	16,9	1,11	0,95	0,06	0,08	0,79	0,07	0,00	-0,32	-0,26	0,52	-0,25	0,00	-0,36	-0,30	0,46	-0,30	0,00
CL29	0,75	0,19	0,43	0,42	0,35	4	23,6	18,4	0,97	0,78	0,12	0,08	0,05	0,75	0,00	-0,23	-0,25	-0,17	0,42	0,00	-0,29	-0,30	-0,21	0,35	0,00
CL30	0,74	0,19	0,44	0,40	0,33	1	23,6	18,6	0,94	0,75	0,74	0,05	0,06	0,14	0,00	0,40	-0,21	-0,24	-0,20	0,00	0,33	-0,25	-0,28	-0,26	0,00
CL31	0,89	0,10	0,32	0,36	0,30	2	23,0	17,0	0,60	0,50	0,02	0,89	0,05	0,04	0,00	-0,19	0,36	-0,19	-0,22	0,00	-0,21	0,30	-0,23	-0,25	0,00
CL32	0,63	0,23	0,48	0,32	0,24	3	23,6	20,0	0,83	0,60	0,02	0,07	0,63	0,28	0,00	-0,19	-0,22	0,32	-0,16	0,00	-0,22	-0,26	0,24	-0,24	0,00
CL33	0,50	0,25	0,50	0,43	0,35	2	24,6	20,0	1,15	0,90	0,42	0,50	0,05	0,04	0,00	-0,31	0,43	-0,13	-0,17	0,00	-0,39	0,35	-0,17	-0,21	0,00
CL34	0,81	0,15	0,39	0,33	0,27	3	23,2	18,6	0,69	0,54	0,11	0,06	0,81	0,02	0,00	-0,25	-0,16	0,33	-0,11	0,00	-0,30	-0,20	0,27	-0,13	0,00

ANEXO 20. *Descripción de los ítems online de Primaria atendiendo a la Teoría Clásica de los Test*

Ítem Online	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial punt.corrg.	Clave	Media p	Media q	Cov(i,X) (Six)	Cov(i,X) (Six)	% Alter. 1	% Alter. 2	% Alter. 3	% Alter. 4	% otras	rbp Alter. 1	rbp Alter. 2	rbp Alter. 3	rbp Alter. 4	rbp otras	rbp (Spur) Alter. 1	rbp (Spur) Alter. 2	rbp (Spur) Alter. 3	rbp (Spur) Alter. 4	rbp (Spur) otras
CL1	0,58	0,24	0,49	0,37	0,30	2	22,2	17,9	1,05	0,80	0,29	0,58	0,03	0,10	0,00	-0,27	0,37	-0,16	-0,11	0,00	-0,34	0,30	-0,19	-0,16	0,00
CL2	0,83	0,14	0,38	0,34	0,28	4	21,3	16,2	0,73	0,59	0,04	0,09	0,05	0,83	0,00	-0,19	-0,18	-0,20	0,34	0,00	-0,22	-0,22	-0,24	0,28	0,00
CL3	0,85	0,12	0,35	0,45	0,40	3	21,5	14,2	0,91	0,78	0,02	0,08	0,85	0,04	0,00	-0,25	-0,34	0,45	-0,14	0,00	-0,27	-0,38	0,40	-0,17	0,00
CL4	0,31	0,21	0,46	0,36	0,29	4	23,5	19,1	0,95	0,74	0,30	0,25	0,15	0,31	0,00	-0,20	-0,12	-0,07	0,36	0,00	-0,28	-0,19	-0,13	0,29	0,00
CL5	0,39	0,24	0,49	0,21	0,12	1	21,9	19,5	0,58	0,34	0,39	0,28	0,10	0,22	0,00	0,21	-0,12	-0,15	0,00	0,00	0,12	-0,20	-0,20	-0,07	0,00
CL6	0,63	0,23	0,48	0,43	0,35	2	22,3	17,3	1,16	0,93	0,13	0,63	0,18	0,05	0,00	-0,29	0,43	-0,14	-0,24	0,00	-0,34	0,35	-0,20	-0,28	0,00
CL7	0,80	0,16	0,40	0,40	0,33	1	21,6	16,0	0,90	0,74	0,80	0,12	0,07	0,01	0,00	0,40	-0,32	-0,16	-0,14	0,00	0,33	-0,37	-0,20	-0,16	0,00
CL8	0,56	0,25	0,50	0,40	0,33	2	22,5	17,9	1,14	0,89	0,21	0,56	0,18	0,06	0,00	-0,19	0,40	-0,28	-0,07	0,00	-0,26	0,33	-0,34	-0,11	0,00
CL9	0,87	0,11	0,34	0,37	0,31	4	21,3	15,1	0,71	0,59	0,09	0,04	0,01	0,87	0,00	-0,30	-0,14	-0,14	0,37	0,00	-0,35	-0,17	-0,16	0,31	0,00
CL10	0,34	0,23	0,47	0,33	0,25	3	23,0	19,1	0,88	0,66	0,36	0,14	0,34	0,15	0,00	-0,08	-0,23	0,33	-0,10	0,00	-0,16	-0,28	0,25	-0,17	0,00
CL11	0,27	0,20	0,44	0,12	0,05	1	21,6	20,0	0,31	0,12	0,27	0,63	0,05	0,05	0,00	0,12	0,07	-0,18	-0,24	0,00	0,05	-0,01	-0,22	-0,27	0,00
CL12	0,46	0,25	0,50	0,33	0,25	4	22,5	18,7	0,95	0,70	0,20	0,16	0,18	0,46	0,00	-0,28	-0,12	-0,03	0,33	0,00	-0,34	-0,18	-0,10	0,25	0,00
CL13	0,32	0,22	0,46	0,26	0,19	3	22,6	19,4	0,70	0,48	0,25	0,26	0,32	0,18	0,00	-0,09	0,00	0,26	-0,22	0,00	-0,17	-0,08	0,19	-0,28	0,00
CL14	0,41	0,24	0,49	0,39	0,31	3	23,1	18,6	1,08	0,84	0,36	0,13	0,41	0,09	0,00	-0,10	-0,23	0,39	-0,21	0,00	-0,19	-0,29	0,31	-0,26	0,00
CL15	0,82	0,15	0,39	0,40	0,34	2	21,5	15,7	0,88	0,73	0,06	0,82	0,09	0,04	0,00	-0,26	0,40	-0,22	-0,17	0,00	-0,30	0,34	-0,26	-0,20	0,00
CL16	0,70	0,21	0,46	0,43	0,36	3	22,0	16,7	1,11	0,90	0,16	0,10	0,70	0,04	0,00	-0,23	-0,24	0,43	-0,19	0,00	-0,29	-0,29	0,36	-0,22	0,00
CL17	0,46	0,25	0,50	0,48	0,41	4	23,4	17,9	1,35	1,10	0,16	0,23	0,14	0,46	0,00	-0,18	-0,21	-0,24	0,48	0,00	-0,25	-0,28	-0,29	0,41	0,00
CL18	0,55	0,25	0,50	0,40	0,32	2	22,5	18,0	1,13	0,88	0,23	0,55	0,12	0,11	0,00	-0,32	0,40	-0,16	-0,05	0,00	-0,38	0,32	-0,21	-0,10	0,00
CL19	0,34	0,23	0,47	0,22	0,14	2	22,2	19,5	0,60	0,38	0,22	0,34	0,06	0,38	0,00	-0,24	0,22	-0,24	0,09	0,00	-0,30	0,14	-0,28	0,01	0,00
CL20	0,27	0,19	0,44	0,31	0,24	4	23,4	19,4	0,77	0,58	0,28	0,18	0,28	0,27	0,00	-0,07	-0,19	-0,08	0,31	0,00	-0,15	-0,25	-0,15	0,24	0,00
CL21	0,54	0,25	0,50	0,28	0,20	1	21,9	18,7	0,79	0,54	0,54	0,05	0,01	0,41	0,00	0,28	-0,25	-0,18	-0,14	0,00	0,20	-0,29	-0,20	-0,22	0,00
CL22	0,73	0,20	0,44	0,50	0,44	3	22,2	15,7	1,26	1,07	0,08	0,12	0,73	0,07	0,00	-0,27	-0,30	0,50	-0,21	0,00	-0,31	-0,35	0,44	-0,25	0,00
CL23	0,87	0,12	0,34	0,49	0,44	2	21,5	13,3	0,95	0,83	0,05	0,87	0,05	0,03	0,00	-0,29	0,49	-0,29	-0,23	0,00	-0,33	0,44	-0,32	-0,26	0,00
CL24	0,74	0,19	0,44	0,29	0,22	4	21,4	17,6	0,73	0,54	0,07	0,17	0,02	0,74	0,00	-0,29	-0,07	-0,21	0,29	0,00	-0,33	-0,14	-0,23	0,22	0,00
CL25	0,45	0,25	0,50	0,31	0,23	1	22,4	18,8	0,89	0,64	0,45	0,02	0,48	0,05	0,00	0,31	-0,20	-0,17	-0,20	0,00	0,23	-0,23	-0,25	-0,24	0,00
CL26	0,70	0,21	0,46	0,45	0,38	2	22,1	16,5	1,17	0,96	0,08	0,70	0,06	0,16	0,00	-0,25	0,45	-0,30	-0,18	0,00	-0,29	0,38	-0,34	-0,24	0,00
CL27	0,83	0,14	0,37	0,49	0,43	1	21,7	14,3	1,03	0,89	0,83	0,06	0,06	0,04	0,00	0,49	-0,30	-0,26	-0,23	0,00	0,43	-0,34	-0,30	-0,26	0,00
CL28	0,70	0,21	0,46	0,56	0,50	3	22,5	15,6	1,44	1,24	0,11	0,11	0,70	0,09	0,00	-0,34	-0,27	0,56	-0,23	0,00	-0,39	-0,32	0,50	-0,28	0,00
CL29	0,67	0,22	0,47	0,46	0,39	4	22,3	16,8	1,23	1,00	0,17	0,09	0,07	0,67	0,00	-0,23	-0,22	-0,26	0,46	0,00	-0,29	-0,27	-0,30	0,39	0,00
CL30	0,66	0,22	0,47	0,42	0,34	1	22,1	17,1	1,12	0,90	0,66	0,10	0,07	0,17	0,00	0,42	-0,33	-0,25	-0,10	0,00	0,34	-0,37	-0,29	-0,16	0,00
CL31	0,87	0,11	0,33	0,40	0,35	2	21,3	14,5	0,75	0,64	0,03	0,87	0,04	0,06	0,00	-0,25	0,40	-0,22	-0,20	0,00	-0,27	0,35	-0,25	-0,24	0,00
CL32	0,60	0,24	0,49	0,31	0,23	3	21,9	18,3	0,87	0,63	0,02	0,09	0,60	0,29	0,00	-0,17	-0,25	0,31	-0,13	0,00	-0,19	-0,29	0,23	-0,21	0,00
CL33	0,49	0,25	0,50	0,47	0,39	2	23,1	17,8	1,32	1,07	0,43	0,49	0,04	0,03	0,00	-0,32	0,47	-0,23	-0,15	0,00	-0,39	0,39	-0,26	-0,18	0,00
CL34	0,83	0,14	0,38	0,38	0,32	3	21,4	15,7	0,82	0,67	0,09	0,07	0,83	0,01	0,00	-0,30	-0,18	0,38	-0,12	0,00	-0,35	-0,22	0,32	-0,13	0,00

ANEXO 21. *Descripción de los ítems en papel de Secundaria atendiendo a la Teoría Clásica de los Test*

Ítem Papel	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial p.corregida	Clave	Media p	Media q	Cov(i,X) (Six)	Cov(i,X) (Six)	% Alter. 1	% Alter. 2	% Alter. 3	% Alter. 4	% otras	rbp Alter.1	rbp Alter.2	rbp Alter. 3	rbp Alter. 4	rbp otras	rbp (Spur) Alter. 1	rbp (Spur) Alter. 2	rbp (Spur) Alter. 3	rbp (Spur) Alter. 4	rbp (Spur) otras
CL1	0,72	0,20	0,45	0,21	0,13	3	24,5	22,1	0,49	0,29	0,15	0,12	0,72	0,01	0,00	-0,11	-0,13	0,21	-0,10	0,00	-0,18	-0,19	0,13	-0,12	0,00
CL2	0,88	0,10	0,32	0,36	0,31	1	24,5	18,7	0,61	0,50	0,88	0,07	0,03	0,02	0,00	0,36	-0,29	-0,18	-0,10	0,00	0,31	-0,34	-0,21	-0,13	0,00
CL3	0,83	0,14	0,37	0,32	0,25	4	24,6	20,2	0,61	0,47	0,01	0,13	0,03	0,83	0,00	-0,22	-0,15	-0,28	0,32	0,00	-0,24	-0,21	-0,31	0,25	0,00
CL4	0,79	0,16	0,40	0,38	0,31	4	24,8	20,0	0,80	0,63	0,10	0,06	0,05	0,79	0,00	-0,20	-0,24	-0,18	0,38	0,00	-0,25	-0,28	-0,22	0,31	0,00
CL5	0,88	0,10	0,32	0,39	0,34	2	24,6	18,3	0,65	0,55	0,05	0,88	0,05	0,03	0,00	-0,25	0,39	-0,18	-0,23	0,00	-0,28	0,34	-0,22	-0,26	0,00
CL6	0,72	0,20	0,45	0,25	0,17	1	24,7	21,7	0,59	0,39	0,72	0,17	0,07	0,04	0,00	0,25	-0,02	-0,29	-0,17	0,00	0,17	-0,09	-0,33	-0,20	0,00
CL7	0,59	0,24	0,49	0,32	0,23	3	25,2	21,9	0,82	0,58	0,06	0,13	0,59	0,23	0,00	-0,20	-0,12	0,32	-0,17	0,00	-0,24	-0,19	0,23	-0,24	0,00
CL8	0,49	0,25	0,50	0,35	0,26	1	25,7	22,1	0,89	0,64	0,49	0,31	0,09	0,12	0,00	0,35	-0,17	-0,14	-0,17	0,00	0,26	-0,26	-0,19	-0,23	0,00
CL9	0,39	0,24	0,49	0,22	0,13	3	25,2	22,9	0,56	0,32	0,20	0,29	0,39	0,11	0,00	-0,07	-0,09	0,22	-0,12	0,00	-0,15	-0,18	0,13	-0,18	0,00
CL10	0,54	0,25	0,50	0,33	0,24	4	25,4	22,0	0,84	0,59	0,04	0,16	0,25	0,54	0,00	-0,27	-0,12	-0,15	0,33	0,00	-0,31	-0,19	-0,23	0,24	0,00
CL11	0,86	0,12	0,35	0,36	0,30	2	24,6	19,3	0,64	0,52	0,08	0,86	0,03	0,03	0,00	-0,17	0,36	-0,22	-0,23	0,00	-0,22	0,30	-0,25	-0,27	0,00
CL12	0,78	0,17	0,41	0,40	0,33	4	24,9	19,9	0,85	0,68	0,05	0,09	0,07	0,78	0,00	-0,19	-0,23	-0,23	0,40	0,00	-0,23	-0,28	-0,27	0,33	0,00
CL13	0,88	0,11	0,33	0,48	0,42	2	24,8	17,3	0,81	0,70	0,04	0,88	0,06	0,02	0,00	-0,23	0,48	-0,33	-0,21	0,00	-0,27	0,42	-0,37	-0,24	0,00
CL14	0,64	0,23	0,48	0,45	0,37	1	25,6	20,7	1,11	0,88	0,64	0,11	0,13	0,12	0,00	0,45	-0,25	-0,28	-0,13	0,00	0,37	-0,31	-0,34	-0,19	0,00
CL15	0,59	0,24	0,49	0,41	0,32	4	25,6	21,3	1,03	0,79	0,17	0,08	0,15	0,59	0,00	-0,14	-0,26	-0,21	0,41	0,00	-0,21	-0,31	-0,27	0,32	0,00
CL16	0,67	0,22	0,47	0,33	0,24	3	25,0	21,4	0,79	0,57	0,14	0,13	0,67	0,05	0,00	-0,09	-0,23	0,33	-0,20	0,00	-0,15	-0,29	0,24	-0,24	0,00
CL17	0,45	0,25	0,50	0,39	0,31	4	26,1	22,0	1,01	0,76	0,06	0,38	0,10	0,45	0,00	-0,28	-0,13	-0,21	0,39	0,00	-0,33	-0,22	-0,27	0,31	0,00
CL18	0,80	0,16	0,40	0,46	0,40	1	25,0	19,0	0,94	0,78	0,80	0,03	0,11	0,05	0,00	0,46	-0,21	-0,29	-0,25	0,00	0,40	-0,24	-0,34	-0,29	0,00
CL19	0,35	0,23	0,48	0,22	0,13	4	25,4	23,0	0,54	0,31	0,11	0,38	0,16	0,35	0,00	-0,21	0,06	-0,18	0,22	0,00	-0,27	-0,04	-0,24	0,13	0,00
CL20	0,62	0,24	0,48	0,45	0,37	2	25,7	20,8	1,13	0,90	0,30	0,62	0,05	0,03	0,00	-0,28	0,45	-0,22	-0,27	0,00	-0,36	0,37	-0,26	-0,30	0,00
CL21	0,85	0,13	0,36	0,49	0,43	4	24,9	17,9	0,89	0,77	0,05	0,03	0,08	0,85	0,00	-0,29	-0,29	-0,25	0,49	0,00	-0,33	-0,32	-0,30	0,43	0,00
CL22	0,66	0,22	0,47	0,37	0,29	2	25,2	21,1	0,91	0,69	0,09	0,66	0,05	0,20	0,00	-0,24	0,37	-0,20	-0,17	0,00	-0,29	0,29	-0,24	-0,24	0,00
CL23	0,88	0,11	0,33	0,48	0,43	4	24,8	17,2	0,81	0,70	0,05	0,02	0,05	0,88	0,00	-0,27	-0,28	-0,26	0,48	0,00	-0,31	-0,31	-0,30	0,43	0,00
CL24	0,88	0,10	0,32	0,36	0,31	3	24,5	18,7	0,60	0,50	0,02	0,07	0,88	0,02	0,00	-0,28	-0,13	0,36	-0,28	0,00	-0,31	-0,18	0,31	-0,31	0,00
CL25	0,64	0,23	0,48	0,35	0,27	4	25,2	21,4	0,87	0,64	0,02	0,28	0,06	0,64	0,00	-0,25	-0,15	-0,30	0,35	0,00	-0,27	-0,23	-0,34	0,27	0,00
CL26	0,83	0,14	0,37	0,48	0,42	1	24,9	18,3	0,91	0,78	0,83	0,03	0,03	0,10	0,00	0,48	-0,31	-0,27	-0,25	0,00	0,42	-0,34	-0,31	-0,30	0,00
CL27	0,83	0,14	0,38	0,49	0,43	2	25,0	18,4	0,95	0,81	0,11	0,83	0,02	0,04	0,00	-0,35	0,49	-0,25	-0,20	0,00	-0,40	0,43	-0,27	-0,24	0,00
CL28	0,78	0,17	0,41	0,45	0,38	3	25,1	19,4	0,96	0,79	0,02	0,05	0,78	0,15	0,00	-0,28	-0,31	0,45	-0,23	0,00	-0,30	-0,35	0,38	-0,29	0,00
CL29	0,73	0,20	0,44	0,42	0,34	1	25,1	20,3	0,95	0,76	0,73	0,07	0,03	0,17	0,00	0,42	-0,29	-0,26	-0,19	0,00	0,34	-0,33	-0,28	-0,26	0,00
CL30	0,92	0,07	0,27	0,46	0,42	4	24,5	15,8	0,64	0,57	0,02	0,02	0,04	0,92	0,00	-0,28	-0,24	-0,26	0,46	0,00	-0,31	-0,27	-0,29	0,42	0,00
CL31	0,92	0,08	0,28	0,46	0,42	1	24,5	16,0	0,65	0,58	0,92	0,03	0,03	0,03	0,00	0,46	-0,25	-0,31	-0,21	0,00	0,42	-0,28	-0,34	-0,24	0,00
CL32	0,74	0,19	0,44	0,38	0,30	2	25,0	20,6	0,86	0,66	0,04	0,74	0,19	0,04	0,00	-0,23	0,38	-0,24	-0,16	0,00	-0,26	0,30	-0,31	-0,19	0,00
CL33	0,67	0,22	0,47	0,37	0,28	4	25,2	21,1	0,89	0,67	0,05	0,17	0,11	0,67	0,00	-0,19	-0,13	-0,26	0,37	0,00	-0,23	-0,20	-0,32	0,28	0,00

ANEXO 22. *Descripción de los ítems online de Secundaria atendiendo a la Teoría Clásica de los Test*

Ítem Papel	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial p.corregida	Clave	Media p	Media q	Cov(i,X) (Six)	Cov(i,X) (Six) (Spur)	% Alter. 1	% Alter. 2	% Alter. 3	% Alter. 4	% otras	rbp Alter.1	rbp Alter.2	rbp Alter. 3	rbp Alter. 4	rbp otras	rbp (Spur) Alter. 1	rbp (Spur) Alter. 2	rbp (Spur) Alter. 3	rbp (Spur)
CL1	0,71	0,21	0,45	0,14	0,04	3	24,7	23,3	0,29	0,08	0,18	0,11	0,71	0,01	0,00	-0,10	-0,07	0,14	-0,03	0,00	-0,18	-0,14	0,04	-
CL2	0,90	0,09	0,30	0,26	0,19	1	24,7	20,8	0,35	0,26	0,90	0,05	0,03	0,02	0,00	0,26	-0,19	-0,17	-0,04	0,00	0,19	-0,24	-0,21	-
CL3	0,84	0,14	0,37	0,30	0,23	4	24,9	21,2	0,51	0,38	0,01	0,13	0,02	0,84	0,00	-0,18	-0,20	-0,20	0,30	0,00	-0,19	-0,27	-0,24	
CL4	0,83	0,14	0,38	0,35	0,27	4	25,1	20,8	0,60	0,46	0,08	0,05	0,04	0,83	0,00	-0,20	-0,24	-0,12	0,35	0,00	-0,26	-0,28	-0,17	
CL5	0,90	0,09	0,29	0,34	0,28	2	24,8	19,5	0,46	0,37	0,04	0,90	0,04	0,02	0,00	-0,24	0,34	-0,17	-0,16	0,00	-0,28	0,28	-0,21	-
CL6	0,71	0,21	0,45	0,20	0,11	1	24,9	22,9	0,42	0,22	0,71	0,19	0,06	0,05	0,00	0,20	0,01	-0,25	-0,18	0,00	0,11	-0,08	-0,30	-
CL7	0,58	0,24	0,49	0,29	0,18	3	25,5	22,8	0,65	0,41	0,06	0,14	0,58	0,22	0,00	-0,17	-0,07	0,29	-0,19	0,00	-0,22	-0,14	0,18	-
CL8	0,51	0,25	0,50	0,36	0,26	1	26,0	22,6	0,83	0,58	0,51	0,31	0,09	0,09	0,00	0,36	-0,21	-0,14	-0,15	0,00	0,26	-0,31	-0,20	-
CL9	0,39	0,24	0,49	0,22	0,12	3	25,6	23,5	0,49	0,25	0,21	0,28	0,39	0,12	0,00	-0,04	-0,11	0,22	-0,14	0,00	-0,13	-0,20	0,12	-
CL10	0,55	0,25	0,50	0,30	0,20	4	25,6	22,8	0,70	0,45	0,04	0,15	0,26	0,55	0,00	-0,17	-0,12	-0,17	0,30	0,00	-0,22	-0,19	-0,26	
CL11	0,88	0,10	0,32	0,33	0,26	2	24,9	20,2	0,48	0,38	0,07	0,88	0,02	0,03	0,00	-0,18	0,33	-0,21	-0,19	0,00	-0,23	0,26	-0,24	-
CL12	0,80	0,16	0,40	0,40	0,32	4	25,2	20,6	0,73	0,57	0,05	0,07	0,07	0,80	0,00	-0,15	-0,25	-0,24	0,40	0,00	-0,19	-0,30	-0,29	
CL13	0,89	0,10	0,31	0,41	0,35	2	25,0	19,0	0,60	0,50	0,03	0,89	0,05	0,03	0,00	-0,23	0,41	-0,23	-0,23	0,00	-0,26	0,35	-0,28	-
CL14	0,64	0,23	0,48	0,43	0,34	1	25,8	21,7	0,95	0,72	0,64	0,13	0,11	0,12	0,00	0,43	-0,29	-0,25	-0,10	0,00	0,34	-0,35	-0,31	-
CL15	0,57	0,24	0,49	0,36	0,26	4	25,8	22,4	0,83	0,58	0,18	0,07	0,18	0,57	0,00	-0,15	-0,22	-0,17	0,36	0,00	-0,23	-0,28	-0,25	
CL16	0,70	0,21	0,46	0,27	0,18	3	25,2	22,4	0,58	0,36	0,15	0,10	0,70	0,05	0,00	-0,11	-0,17	0,27	-0,17	0,00	-0,18	-0,23	0,18	-
CL17	0,49	0,25	0,50	0,34	0,23	4	25,9	22,8	0,77	0,52	0,05	0,36	0,11	0,49	0,00	-0,26	-0,10	-0,21	0,34	0,00	-0,30	-0,20	-0,27	
CL18	0,82	0,15	0,39	0,43	0,35	1	25,3	20,2	0,76	0,61	0,82	0,04	0,10	0,05	0,00	0,43	-0,20	-0,24	-0,25	0,00	0,35	-0,24	-0,30	-
CL19	0,37	0,23	0,48	0,24	0,14	4	25,8	23,5	0,53	0,30	0,11	0,36	0,17	0,37	0,00	-0,22	0,03	-0,16	0,24	0,00	-0,28	-0,07	-0,24	
CL20	0,71	0,20	0,45	0,41	0,32	2	25,5	21,4	0,85	0,64	0,24	0,71	0,04	0,01	0,00	-0,29	0,41	-0,23	-0,15	0,00	-0,37	0,32	-0,27	-
CL21	0,87	0,12	0,34	0,48	0,42	4	25,2	18,7	0,76	0,64	0,04	0,02	0,08	0,87	0,00	-0,26	-0,24	-0,31	0,48	0,00	-0,30	-0,27	-0,36	
CL22	0,67	0,22	0,47	0,38	0,29	2	25,6	21,9	0,82	0,60	0,08	0,67	0,04	0,21	0,00	-0,23	0,38	-0,17	-0,20	0,00	-0,28	0,29	-0,21	-
CL23	0,89	0,10	0,31	0,42	0,36	4	25,0	18,8	0,59	0,50	0,05	0,02	0,04	0,89	0,00	-0,27	-0,22	-0,21	0,42	0,00	-0,31	-0,25	-0,25	
CL24	0,90	0,09	0,30	0,25	0,19	3	24,7	20,8	0,34	0,25	0,01	0,07	0,90	0,02	0,00	-0,18	-0,10	0,25	-0,22	0,00	-0,20	-0,16	0,19	-
CL25	0,67	0,22	0,47	0,35	0,26	4	25,5	22,0	0,76	0,54	0,01	0,27	0,05	0,67	0,00	-0,18	-0,21	-0,25	0,35	0,00	-0,20	-0,30	-0,29	
CL26	0,86	0,12	0,35	0,44	0,38	1	25,2	19,4	0,71	0,58	0,86	0,02	0,03	0,09	0,00	0,44	-0,25	-0,28	-0,25	0,00	0,38	-0,28	-0,31	-
CL27	0,86	0,12	0,34	0,42	0,36	2	25,1	19,5	0,66	0,55	0,08	0,86	0,03	0,03	0,00	-0,28	0,42	-0,25	-0,17	0,00	-0,33	0,36	-0,28	-
CL28	0,81	0,15	0,39	0,42	0,34	3	25,3	20,4	0,76	0,60	0,02	0,05	0,81	0,12	0,00	-0,27	-0,26	0,42	-0,22	0,00	-0,30	-0,30	0,34	-
CL29	0,78	0,17	0,42	0,40	0,32	1	25,3	20,9	0,76	0,59	0,78	0,05	0,03	0,15	0,00	0,40	-0,20	-0,28	-0,22	0,00	0,32	-0,24	-0,31	-
CL30	0,93	0,06	0,25	0,40	0,35	4	24,8	17,6	0,47	0,41	0,03	0,02	0,03	0,93	0,00	-0,23	-0,24	-0,22	0,40	0,00	-0,26	-0,27	-0,25	
CL31	0,92	0,08	0,28	0,45	0,40	1	25,0	17,5	0,56	0,49	0,92	0,04	0,02	0,02	0,00	0,45	-0,29	-0,22	-0,24	0,00	0,40	-0,32	-0,25	-
CL32	0,74	0,19	0,44	0,34	0,25	2	25,2	21,7	0,68	0,48	0,04	0,74	0,19	0,03	0,00	-0,23	0,34	-0,21	-0,11	0,00	-0,27	0,25	-0,29	-
CL33	0,67	0,22	0,47	0,38	0,29	4	25,6	21,9	0,82	0,60	0,05	0,20	0,09	0,67	0,00	-0,20	-0,21	-0,19	0,38	0,00	-0,25	-0,29	-0,25	

ANEXO 23. *Parámetros y ajuste del modelo de 1 Parámetro de la prueba en papel de Primaria*

Modelo 1 Parámetro - Papel

	b	se(b)	Chi cuadrado	gl	p
Ítem1	-0,875	0,077	4,604	8,000	0,799
Ítem2	-2,072	0,098	6,969	8,000	0,540
Ítem3	-2,434	0,108	29,665	8,000	0,000
Ítem4	0,260	0,073	20,334	8,000	0,009
Ítem5	0,362	0,073	25,086	8,000	0,002
Ítem6	-0,880	0,077	7,299	8,000	0,505
Ítem7	-2,030	0,096	21,841	8,000	0,005
Ítem8	-0,574	0,075	11,845	8,000	0,158
Ítem9	-2,605	0,115	14,193	8,000	0,077
Ítem10	0,264	0,073	14,113	8,000	0,079
Ítem11	0,611	0,075	37,267	8,000	0,000
Ítem12	0,031	0,073	12,200	8,000	0,142
Ítem13	0,465	0,074	23,746	8,000	0,003
Ítem14	0,023	0,073	19,626	8,000	0,012
Ítem15	-1,739	0,090	18,198	8,000	0,020
Ítem16	-1,469	0,085	17,809	8,000	0,023
Ítem17	-0,249	0,073	17,459	8,000	0,026
Ítem18	-0,528	0,074	5,698	8,000	0,681
Ítem19	0,388	0,074	23,347	8,000	0,003
Ítem20	0,629	0,075	42,564	8,000	0,000
Ítem21	-0,144	0,073	24,673	8,000	0,002
Ítem22	-1,623	0,087	39,464	8,000	0,000
Ítem23	-2,072	0,098	23,620	8,000	0,003
Ítem24	-0,965	0,078	7,774	8,000	0,456
Ítem25	-0,060	0,073	7,047	8,000	0,532
Ítem26	-1,308	0,082	42,045	8,000	0,000
Ítem27	-2,132	0,099	39,191	8,000	0,000
Ítem28	-1,475	0,085	67,286	8,000	0,000
Ítem29	-1,261	0,081	12,263	8,000	0,140
Ítem30	-1,145	0,080	12,470	8,000	0,131
Ítem31	-2,233	0,102	16,331	8,000	0,038
Ítem32	-0,533	0,074	17,551	8,000	0,025
Ítem33	0,001	0,073	16,494	8,000	0,036
Ítem34	-1,584	0,087	5,601	8,000	0,692

ANEXO 24. *Parámetros y ajuste del modelo de 2 Parámetro de la prueba en papel de Primaria*

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,805	-1,071	0,090	0,131	0,008	5,660	7,000	0,580
Ítem2	0,828	-2,463	0,113	0,291	0,030	5,074	7,000	0,651
Ítem3	1,728	-1,791	0,183	0,122	0,018	5,003	7,000	0,660
Ítem4	0,706	0,339	0,083	0,102	-0,003	22,888	7,000	0,002
Ítem5	0,392	0,804	0,073	0,216	-0,010	9,771	7,000	0,202
Ítem6	0,795	-1,087	0,090	0,134	0,009	12,978	7,000	0,073
Ítem7	0,872	-2,317	0,114	0,259	0,027	17,565	7,000	0,014
Ítem8	0,763	-0,736	0,086	0,114	0,006	15,499	7,000	0,030
Ítem9	1,008	-2,661	0,140	0,298	0,039	7,987	7,000	0,334
Ítem10	0,553	0,425	0,078	0,130	-0,004	15,079	7,000	0,035
Ítem11	0,365	1,465	0,074	0,334	-0,021	12,009	7,000	0,100
Ítem12	0,753	0,030	0,084	0,091	0,000	14,797	7,000	0,039
Ítem13	0,774	0,572	0,087	0,104	-0,004	23,951	7,000	0,001
Ítem14	0,762	0,019	0,084	0,091	0,000	18,261	7,000	0,011
Ítem15	1,236	-1,559	0,125	0,128	0,013	6,365	7,000	0,498
Ítem16	1,107	-1,419	0,112	0,126	0,011	6,032	7,000	0,536
Ítem17	0,925	-0,282	0,092	0,081	0,002	13,042	7,000	0,071
Ítem18	0,887	-0,604	0,091	0,094	0,004	9,662	7,000	0,209
Ítem19	0,462	0,742	0,075	0,179	-0,008	13,010	7,000	0,072
Ítem20	0,842	0,730	0,090	0,106	-0,005	3,961	7,000	1,497
Ítem21	0,375	-0,353	0,072	0,180	0,005	8,116	7,000	0,322
Ítem22	1,421	-1,343	0,134	0,101	0,010	7,088	7,000	0,420
Ítem23	1,324	-1,773	0,139	0,141	0,016	10,779	7,000	0,149
Ítem24	0,657	-1,391	0,085	0,187	0,013	7,528	7,000	0,376
Ítem25	0,635	-0,100	0,080	0,106	0,001	8,691	7,000	0,276
Ítem26	1,388	-1,101	0,126	0,089	0,007	11,060	7,000	0,136
Ítem27	1,730	-1,579	0,172	0,105	0,014	12,885	7,000	0,075
Ítem28	2,102	-1,020	0,187	0,067	0,007	5,348	7,000	0,618
Ítem29	1,077	-1,244	0,107	0,115	0,009	12,266	7,000	0,092
Ítem30	0,857	-1,331	0,095	0,145	0,011	8,848	7,000	0,264
Ítem31	1,254	-1,972	0,140	0,166	0,020	5,039	7,000	0,655
Ítem32	0,495	-0,985	0,077	0,191	0,010	5,954	7,000	0,545
Ítem33	0,980	-0,008	0,094	0,075	0,000	17,348	7,000	0,015
Ítem34	0,933	-1,722	0,106	0,172	0,015	11,602	7,000	0,114

ANEXO 25. *Parámetros y ajuste del modelo de 3 Parámetro de la prueba en papel de Primaria*

Modelo 3 Parámetros - Papel

	a	b	c	se(a)	se(b)	se(c)	cov(a,b)	cov(a,c)	cov(b,c)	Chi cuadrado	gl	p
Ítem1	0,896	-0,699	0,131	0,230	0,682	0,232	0,145	0,048	0,156	3,367	6,000	0,762
Ítem2	0,800	-2,534	0,000	0,107	0,296	0,003	0,029	0,000	0,000	3,385	6,000	0,759
Ítem3	1,730	-1,657	0,148	0,383	0,546	0,355	0,194	0,117	0,188	4,129	6,000	0,659
Ítem4	1,307	0,861	0,208	0,283	0,147	0,050	0,014	0,011	0,005	9,656	6,000	0,140
Ítem5	0,400	0,875	0,013	0,193	1,445	0,283	0,242	0,050	0,404	6,965	6,000	0,324
Ítem6	0,773	-1,103	0,000	0,087	0,140	0,009	0,009	0,000	0,000	9,572	6,000	0,144
Ítem7	0,839	-2,389	0,000	0,108	0,275	0,034	0,026	0,000	0,002	20,569	6,000	0,002
Ítem8	1,067	0,064	0,259	0,247	0,320	0,100	0,064	0,020	0,031	17,947	6,000	0,006
Ítem9	0,967	-2,756	0,000	0,130	0,303	0,004	0,036	0,000	0,000	0,831	6,000	0,217
Ítem10	1,347	1,141	0,270	0,341	0,158	0,046	0,005	0,012	0,004	12,198	6,000	0,058
Ítem11	1,771	1,690	0,294	0,714	0,207	0,031	-0,091	0,016	-0,001	15,034	6,000	0,020
Ítem12	0,875	0,320	0,094	0,248	0,416	0,140	0,089	0,031	0,057	13,019	6,000	0,043
Ítem13	1,181	0,924	0,148	0,252	0,147	0,051	0,010	0,010	0,005	11,417	6,000	0,076
Ítem14	1,580	0,703	0,260	0,341	0,131	0,047	0,019	0,012	0,005	11,218	6,000	0,082
Ítem15	1,195	-1,590	0,000	0,118	0,132	0,000	0,012	0,000	0,000	6,837	6,000	0,336
Ítem16	1,065	-1,448	0,000	0,107	0,135	0,014	0,011	0,000	0,000	10,472	6,000	0,106
Ítem17	1,416	0,295	0,219	0,283	0,174	0,066	0,035	0,015	0,010	5,809	6,000	0,445
Ítem18	1,233	0,029	0,229	0,291	0,294	0,105	0,073	0,026	0,030	11,178	6,000	0,083
Ítem19	1,610	1,408	0,302	0,556	0,159	0,039	-0,023	0,017	0,001	6,661	6,000	0,353
Ítem20	2,470	0,989	0,201	0,586	0,084	0,027	-0,001	0,010	0,001	20,927	6,000	0,002
Ítem21	0,386	-0,215	0,024	0,183	2,262	0,409	0,381	0,069	0,923	4,114	6,000	0,661
Ítem22	1,562	-1,033	0,185	0,296	0,349	0,173	0,094	0,043	0,058	3,948	6,000	0,684
Ítem23	1,249	-1,840	0,000	0,129	0,149	0,000	0,016	0,000	0,000	16,068	6,000	0,013
Ítem24	0,653	-1,390	0,001	0,085	0,234	0,046	0,015	0,001	0,006	28,390	6,000	0,829
Ítem25	0,851	0,487	0,172	0,264	0,424	0,130	0,093	0,031	0,053	7,593	6,000	0,269
Ítem26	1,554	-0,800	0,163	0,210	0,188	0,088	0,031	0,012	0,015	21,520	6,000	0,001
Ítem27	1,963	-1,265	0,221	0,290	0,219	0,124	0,052	0,023	0,024	10,790	6,000	0,095
Ítem28	2,014	-1,020	0,000	0,179	0,071	0,006	0,007	0,000	0,000	6,784	6,000	0,341
Ítem29	1,263	-0,761	0,216	0,234	0,348	0,138	0,071	0,026	0,046	20,781	6,000	0,002
Ítem30	0,833	-1,353	0,000	0,092	0,150	0,004	0,011	0,000	0,000	9,155	6,000	0,165
Ítem31	1,200	-2,034	0,000	0,131	0,172	0,000	0,019	0,000	0,000	4,918	6,000	0,554
Ítem32	0,489	-0,990	0,000	0,075	0,196	0,007	0,010	0,000	0,000	5,773	6,000	0,449
Ítem33	1,146	0,225	0,089	0,193	0,175	0,067	0,023	0,010	0,011	11,851	6,000	0,065
Ítem34	0,891	-1,777	0,000	0,100	0,181	0,004	0,015	0,000	0,000	2,978	6,000	0,812

ANEXO 26. *Parámetros y ajuste del modelo de 1 Parámetro de la prueba online de Primaria*

Modelo 1 Parámetro - Online

	b	se(b)	Chi cuadrado	gl	p
Ítem1	-0,388	0,073	2,912	8,000	0,940
Ítem2	-1,822	0,091	11,276	8,000	0,187
Ítem3	-2,031	0,096	4,138	8,000	0,000
Ítem4	0,946	0,077	5,185	8,000	0,738
Ítem5	0,520	0,074	5,393	8,000	0,000
Ítem6	-0,642	0,075	8,683	8,000	0,370
Ítem7	-1,589	0,086	19,735	8,000	0,011
Ítem8	-0,281	0,073	18,511	8,000	0,018
Ítem9	-2,151	0,099	11,490	8,000	0,175
Ítem10	0,755	0,075	2,683	8,000	0,001
Ítem11	1,175	0,080	7,802	8,000	0,000
Ítem12	0,173	0,073	4,728	8,000	0,786
Ítem13	0,902	0,077	5,417	8,000	0,000
Ítem14	0,406	0,073	13,374	8,000	0,100
Ítem15	-1,715	0,089	12,427	8,000	0,133
Ítem16	-0,986	0,078	19,680	8,000	0,012
Ítem17	0,182	0,073	3,458	8,000	0,000
Ítem18	-0,215	0,073	15,315	8,000	0,053
Ítem19	0,760	0,075	3,671	8,000	0,000
Ítem20	1,186	0,080	25,366	8,000	0,001
Ítem21	-0,176	0,073	23,114	8,000	0,003
Ítem22	-1,165	0,080	5,410	8,000	0,000
Ítem23	-2,142	0,099	6,327	8,000	0,000
Ítem24	-1,214	0,080	26,182	8,000	0,001
Ítem25	0,220	0,073	19,610	8,000	0,012
Ítem26	-0,986	0,078	16,093	8,000	0,041
Ítem27	-1,860	0,092	5,448	8,000	0,000
Ítem28	-0,996	0,078	7,322	8,000	0,000
Ítem29	-0,803	0,076	20,642	8,000	0,008
Ítem30	-0,770	0,076	9,557	8,000	0,297
Ítem31	-2,205	0,101	18,594	8,000	0,017
Ítem32	-0,468	0,074	19,930	8,000	0,011
Ítem33	0,045	0,073	21,984	8,000	0,005
Ítem34	-1,808	0,091	12,764	8,000	0,120

ANEXO 27. *Parámetros y ajuste del modelo de 2 Parámetro de la prueba online de Primaria*

Modelo 2 Parámetros - Online

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,755	-0,502	0,084	0,101	0,004	10,892	7,000	0,143
Ítem2	0,928	-1,979	0,114	0,207	0,021	6,004	7,000	0,539
Ítem3	1,644	-1,533	0,166	0,107	0,014	10,374	7,000	0,168
Ítem4	0,789	1,159	0,088	0,140	-0,009	15,220	7,000	0,033
Ítem5	0,294	1,533	0,071	0,416	-0,025	4,502	7,000	0,721
Ítem6	0,942	-0,702	0,094	0,093	0,005	7,406	7,000	0,388
Ítem7	1,108	-1,527	0,117	0,136	0,013	11,672	7,000	0,112
Ítem8	0,843	-0,339	0,087	0,087	0,002	17,240	7,000	0,016
Ítem9	1,238	-1,904	0,142	0,166	0,020	4,626	7,000	0,705
Ítem10	0,637	1,102	0,081	0,162	-0,010	3,248	7,000	0,000
Ítem11	0,104	9,762	0,076	7,121	-0,537	9,051	7,000	0,249
Ítem12	0,603	0,258	0,077	0,114	-0,002	3,798	7,000	0,803
Ítem13	0,498	1,634	0,078	0,269	-0,018	3,832	7,000	2,632
Ítem14	0,839	0,467	0,086	0,093	-0,003	9,170	7,000	0,241
Ítem15	1,220	-1,545	0,127	0,129	0,013	5,367	7,000	0,615
Ítem16	1,016	-1,015	0,101	0,104	0,007	13,216	7,000	0,067
Ítem17	1,137	0,157	0,100	0,069	-0,001	10,835	7,000	0,146
Ítem18	0,774	-0,278	0,084	0,092	0,002	10,968	7,000	0,140
Ítem19	0,337	1,978	0,073	0,453	-0,030	2,961	7,000	0,889
Ítem20	0,674	1,656	0,086	0,211	-0,015	28,610	7,000	0,000
Ítem21	0,463	-0,346	0,074	0,147	0,004	7,687	7,000	0,361
Ítem22	1,476	-0,954	0,131	0,077	0,006	19,501	7,000	0,007
Ítem23	2,350	-1,383	0,242	0,079	0,013	3,542	7,000	0,831
Ítem24	0,575	-1,949	0,087	0,287	0,022	18,444	7,000	0,010
Ítem25	0,564	0,352	0,076	0,124	-0,003	10,905	7,000	0,143
Ítem26	1,198	-0,913	0,111	0,087	0,006	6,028	7,000	0,536
Ítem27	1,936	-1,308	0,185	0,082	0,010	7,237	7,000	0,405
Ítem28	1,954	-0,725	0,164	0,058	0,004	10,987	7,000	0,139
Ítem29	1,179	-0,755	0,107	0,080	0,005	9,176	7,000	0,240
Ítem30	0,948	-0,835	0,095	0,099	0,006	13,743	7,000	0,056
Ítem31	1,454	-1,769	0,158	0,137	0,018	4,555	7,000	0,714
Ítem32	0,516	-0,828	0,077	0,167	0,008	19,055	7,000	0,008
Ítem33	1,084	0,031	0,098	0,070	0,000	11,331	7,000	0,125
Ítem34	1,183	-1,658	0,127	0,143	0,015	4,612	7,000	0,707

ANEXO 28. *Parámetros y ajuste del modelo de 3 Parámetro de la prueba online de Primaria*

Modelo 3 Parámetros -Online

	a	b	c	se(a)	se(b)	se(c)	cov(a,b)	cov(a,c)	cov(b,c)	Chi cuadrado	gl	p
Ítem1	0,736	-0,501	0,000	0,082	0,109	0,010	0,004	0,000	0,000	5,298	6,000	0,506
Ítem2	0,901	-2,026	0,000	0,107	0,210	0,005	0,020	0,000	0,000	8,344	6,000	0,214
Ítem3	1,525	-1,599	0,000	0,151	0,116	0,003	0,014	0,000	0,000	8,557	6,000	0,200
Ítem4	1,040	1,274	0,000	0,249	0,157	0,055	0,005	0,012	0,005	6,245	6,000	0,396
Ítem5	1,046	2,191	0,000	0,564	0,390	0,056	-0,142	0,027	-0,006	7,603	6,000	0,269
Ítem6	0,935	-0,691	0,000	0,092	0,095	0,002	0,005	0,000	0,000	9,543	6,000	0,145
Ítem7	1,798	-0,403	0,000	0,471	0,321	0,106	0,134	0,042	0,032	10,450	6,000	0,107
Ítem8	1,967	0,568	0,000	0,409	0,123	0,043	0,026	0,012	0,004	5,911	6,000	0,433
Ítem9	1,334	-1,447	0,000	0,308	0,686	0,308	0,194	0,080	0,206	10,797	6,000	0,095
Ítem10	2,327	1,330	0,000	0,668	0,105	0,027	-0,013	0,013	0,001	12,079	6,000	0,060
Ítem11	0,116	9,006	0,000	0,086	6,132	0,131	-0,403	0,005	0,185	9,070	6,000	0,170
Ítem12	0,788	0,791	0,000	0,277	0,456	0,139	0,103	0,035	0,061	7,596	6,000	0,269
Ítem13	2,284	1,490	0,000	0,529	0,121	0,021	-0,026	0,006	0,001	15,257	6,000	0,018
Ítem14	1,407	0,886	0,000	0,318	0,137	0,052	0,018	0,013	0,005	8,200	6,000	0,224
Ítem15	1,168	-1,586	0,000	0,149	0,329	0,173	0,040	0,016	0,052	15,522	6,000	0,017
Ítem16	1,428	-0,248	0,000	0,318	0,305	0,111	0,084	0,030	0,033	5,900	6,000	0,434
Ítem17	1,472	0,429	0,000	0,241	0,123	0,050	0,017	0,009	0,005	4,765	6,000	0,574
Ítem18	1,258	0,527	0,000	0,368	0,270	0,093	0,080	0,030	0,024	7,297	6,000	0,294
Ítem19	0,575	2,423	0,000	0,428	0,494	0,158	-0,015	0,064	0,016	2,651	6,000	0,851
Ítem20	1,774	1,549	0,000	0,409	0,135	0,024	-0,028	0,006	0,000	6,010	6,000	0,422
Ítem21	0,636	0,702	0,000	0,317	0,933	0,216	0,265	0,064	0,198	3,120	6,000	0,794
Ítem22	3,128	-0,138	0,000	0,666	0,110	0,049	0,052	0,021	0,004	12,472	6,000	0,052
Ítem23	2,813	-1,109	0,000	0,500	0,172	0,115	0,071	0,039	0,017	4,858	6,000	0,562
Ítem24	0,565	-1,976	0,000	0,083	0,288	0,002	0,021	0,000	0,000	17,501	6,000	0,008
Ítem25	0,939	1,119	0,000	0,333	0,276	0,091	0,053	0,027	0,021	9,409	6,000	0,152
Ítem26	1,220	-0,794	0,000	0,207	0,344	0,160	0,063	0,028	0,053	7,673	6,000	0,263
Ítem27	2,384	-0,912	0,000	0,386	0,175	0,099	0,055	0,026	0,015	9,681	6,000	0,139
Ítem28	2,399	-0,472	0,000	0,324	0,106	0,058	0,025	0,012	0,005	8,252	6,000	0,220
Ítem29	1,493	-0,290	0,000	0,238	0,192	0,081	0,036	0,014	0,014	10,836	6,000	0,094
Ítem30	0,911	-0,845	0,000	0,091	0,105	0,004	0,006	0,000	0,000	6,740	6,000	0,346
Ítem31	1,385	-1,830	0,000	0,145	0,142	0,007	0,017	0,000	0,000	5,280	6,000	0,508
Ítem32	0,518	-0,819	0,000	0,075	0,166	0,002	0,008	0,000	0,000	21,578	6,000	0,001
Ítem33	1,327	0,305	0,000	0,241	0,167	0,069	0,029	0,014	0,010	8,140	6,000	0,228
Ítem34	1,241	-1,277	0,000	0,277	0,631	0,276	0,160	0,065	0,170	2,636	6,000	0,853

ANEXO 29. *Parámetros y ajuste del modelo de 1 Parámetro de la prueba en papel de Secundaria*

Modelo 1 Parámetro - Papel

	b	se(b)	Chi cuadrado	gl	p
Ítem1	-1,176	0,084	17,740	8,000	0,023
Ítem2	-2,511	0,116	16,281	8,000	0,039
Ítem3	-2,201	0,106	8,307	8,000	0,404
Ítem4	-1,594	0,091	16,654	8,000	0,034
Ítem5	-2,359	0,110	58,461	8,000	0,000
Ítem6	-1,116	0,083	38,523	8,000	0,000
Ítem7	-0,593	0,078	32,421	8,000	0,000
Ítem8	-0,029	0,076	16,613	8,000	0,034
Ítem9	0,359	0,076	24,876	8,000	0,002
Ítem10	-0,347	0,076	13,968	8,000	0,083
Ítem11	-2,283	0,108	12,821	8,000	0,118
Ítem12	-1,638	0,092	23,665	8,000	0,003
Ítem13	-2,305	0,109	92,506	8,000	0,000
Ítem14	-0,789	0,079	50,088	8,000	0,000
Ítem15	-0,716	0,079	17,567	8,000	0,025
Ítem16	-0,847	0,080	16,484	8,000	0,036
Ítem17	0,165	0,076	62,902	8,000	0,000
Ítem18	-1,823	0,096	51,356	8,000	0,000
Ítem19	0,631	0,077	36,349	8,000	0,000
Ítem20	-0,800	0,079	48,742	8,000	0,000
Ítem21	-2,201	0,106	53,350	8,000	0,000
Ítem22	-1,030	0,082	21,018	8,000	0,007
Ítem23	-2,475	0,114	53,584	8,000	0,000
Ítem24	-2,393	0,112	55,755	8,000	0,000
Ítem25	-0,875	0,080	32,979	8,000	0,000
Ítem26	-2,002	0,100	31,487	8,000	0,000
Ítem27	-1,823	0,096	93,428	8,000	0,000
Ítem28	-1,559	0,090	33,702	8,000	0,000
Ítem29	-1,249	0,085	24,472	8,000	0,002
Ítem30	-2,811	0,127	57,527	8,000	0,000
Ítem31	-2,667	0,122	197,895	8,000	0,000
Ítem32	-1,146	0,083	39,117	8,000	0,000
Ítem33	-0,902	0,080	67,943	8,000	0,000
Ítem34	-1,176	0,084	17,740	8,000	0,023

ANEXO 30. *Parámetros y ajuste del modelo de 2 Parámetro de la prueba en papel de Secundaria*

Modelo 2 Parámetros - Papel

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,325	-3,195	0,082	0,798	0,063	4,312	7,000	0,743
Ítem2	1,227	-2,208	0,141	0,197	0,024	6,825	7,000	0,447
Ítem3	0,939	-2,342	0,118	0,252	0,027	9,151	7,000	0,242
Ítem4	0,831	-1,872	0,103	0,213	0,019	5,263	7,000	0,628
Ítem5	1,512	-1,827	0,159	0,141	0,018	3,800	7,000	0,803
Ítem6	0,319	-3,089	0,081	0,784	0,061	15,224	7,000	0,033
Ítem7	0,616	-0,910	0,085	0,159	0,009	22,914	7,000	0,002
Ítem8	0,467	-0,087	0,079	0,143	0,001	8,256	7,000	0,311
Ítem9	0,350	0,854	0,077	0,258	-0,014	10,423	7,000	0,166
Ítem10	0,587	-0,566	0,083	0,138	0,006	20,772	7,000	0,004
Ítem11	0,774	-2,827	0,113	0,365	0,039	6,847	7,000	0,445
Ítem12	1,155	-1,508	0,121	0,138	0,013	5,056	7,000	0,653
Ítem13	1,853	-1,603	0,187	0,110	0,015	6,043	7,000	0,535
Ítem14	1,229	-0,702	0,120	0,084	0,006	10,642	7,000	0,155
Ítem15	0,805	-0,873	0,094	0,125	0,008	11,402	7,000	0,122
Ítem16	0,658	-1,218	0,088	0,179	0,012	6,799	7,000	0,450
Ítem17	1,078	0,153	0,110	0,073	-0,001	26,111	7,000	0,000
Ítem18	1,399	-1,481	0,140	0,120	0,013	11,220	7,000	0,129
Ítem19	0,290	1,844	0,078	0,528	-0,037	10,481	7,000	0,163
Ítem20	1,284	-0,691	0,123	0,081	0,005	12,367	7,000	0,089
Ítem21	1,617	-1,642	0,164	0,121	0,015	3,539	7,000	0,831
Ítem22	0,910	-1,133	0,101	0,131	0,010	22,176	7,000	0,002
Ítem23	1,667	-1,815	0,174	0,132	0,018	6,344	7,000	0,500
Ítem24	1,265	-2,064	0,141	0,179	0,022	41,569	7,000	0,000
Ítem25	0,709	-1,179	0,090	0,164	0,011	10,645	7,000	0,155
Ítem26	1,403	-1,622	0,143	0,130	0,015	4,806	7,000	0,684
Ítem27	1,684	-1,333	0,163	0,099	0,012	13,690	7,000	0,057
Ítem28	1,395	-1,271	0,137	0,107	0,011	5,841	7,000	0,558
Ítem29	1,084	-1,206	0,113	0,121	0,010	11,195	7,000	0,130
Ítem30	1,498	-2,182	0,168	0,173	0,024	11,993	7,000	0,101
Ítem31	1,781	-1,889	0,190	0,133	0,020	4,273	7,000	0,748
Ítem32	1,052	-1,132	0,110	0,118	0,010	5,292	7,000	0,624
Ítem33	0,758	-1,150	0,093	0,153	0,011	4,784	7,000	0,686
Ítem34	0,325	-3,195	0,082	0,798	0,063	4,312	7,000	0,743

ANEXO 31. *Parámetros y ajuste del modelo de 3 Parámetro de la prueba en papel de Secundaria*

Modelo 3 Parámetros - Papel

	a	b	c	se(a)	se(b)	se(c)	cov(a,b)	cov(a,c)	cov(b,c)	Chi cuadrado	gl	p
Ítem1	0,318	-3,241	0,005	0,082	1,117	0,134	0,077	0,002	0,101	2,439	6,000	0,875
Ítem2	1,187	-2,256	0,000	0,136	0,206	0,008	0,024	0,000	0,000	7,049	6,000	0,316
Ítem3	0,922	-2,373	0,000	0,114	0,255	0,000	0,026	0,000	0,000	8,597	6,000	0,198
Ítem4	0,825	-1,877	0,000	0,101	0,214	0,000	0,019	0,000	0,000	7,798	6,000	0,253
Ítem5	1,445	-1,869	0,000	0,151	0,149	0,001	0,018	0,000	0,000	3,898	6,000	0,690
Ítem6	0,328	-3,006	0,000	0,080	0,733	0,002	0,056	0,000	0,000	13,825	6,000	0,032
Ítem7	1,318	0,444	0,384	0,374	0,242	0,074	0,060	0,022	0,016	12,800	6,000	0,046
Ítem8	0,665	0,731	0,194	0,262	0,585	0,150	0,118	0,035	0,084	10,330	6,000	0,111
Ítem9	0,384	1,132	0,052	NaN	NaN	NaN	-13,669	-2,652	-18,667	7,336	6,000	0,291
Ítem10	0,585	-0,563	0,000	0,083	0,139	0,001	0,006	0,000	0,000	12,730	6,000	0,048
Ítem11	0,768	-2,758	0,054	0,207	1,731	0,689	0,331	0,119	1,167	4,948	6,000	0,550
Ítem12	1,196	-1,359	0,077	0,183	0,323	0,139	0,051	0,017	0,041	4,514	6,000	0,607
Ítem13	1,816	-1,607	0,000	0,184	0,115	0,000	0,016	0,000	0,000	7,870	6,000	0,248
Ítem14	1,233	-0,689	0,000	0,120	0,085	0,000	0,006	0,000	0,000	10,389	6,000	0,109
Ítem15	0,932	-0,478	0,135	0,232	0,497	0,166	0,103	0,033	0,081	7,065	6,000	0,315
Ítem16	0,664	-1,202	0,000	0,088	0,177	0,000	0,012	0,000	0,000	3,803	6,000	0,703
Ítem17	1,352	0,335	0,084	0,232	0,131	0,055	0,017	0,009	0,006	17,386	6,000	0,008
Ítem18	1,615	-1,147	0,173	0,279	0,276	0,131	0,068	0,027	0,033	7,849	6,000	0,249
Ítem19	0,288	1,903	0,005	0,087	0,727	0,076	-0,019	0,003	0,036	16,424	6,000	0,012
Ítem20	1,696	-0,349	0,152	0,253	0,141	0,067	0,026	0,011	0,008	8,881	6,000	0,180
Ítem21	1,794	-1,398	0,142	0,249	0,199	0,099	0,041	0,013	0,016	5,329	6,000	0,502
Ítem22	1,262	-0,426	0,258	0,256	0,288	0,104	0,061	0,020	0,028	14,794	6,000	0,022
Ítem23	1,630	-1,830	0,000	0,169	0,137	0,000	0,018	0,000	0,000	7,306	6,000	0,293
Ítem24	1,250	-2,076	0,000	0,137	0,181	0,000	0,021	0,000	0,000	29,992	6,000	0,000
Ítem25	1,350	0,090	0,384	0,335	0,264	0,083	0,067	0,021	0,020	6,486	6,000	0,371
Ítem26	1,737	-1,161	0,232	0,280	0,233	0,110	0,055	0,019	0,022	6,055	6,000	0,417
Ítem27	2,010	-1,045	0,154	0,301	0,172	0,090	0,043	0,017	0,013	10,934	6,000	0,090
Ítem28	1,385	-1,264	0,000	0,136	0,109	0,000	0,011	0,000	0,000	6,286	6,000	0,392
Ítem29	1,075	-1,203	0,000	0,112	0,123	0,000	0,010	0,000	0,000	6,438	6,000	0,376
Ítem30	1,456	-2,220	0,000	0,161	0,178	0,000	0,024	0,000	0,000	13,562	6,000	0,035
Ítem31	1,728	-1,915	0,000	0,182	0,139	0,000	0,020	0,000	0,000	5,500	6,000	0,481
Ítem32	1,037	-1,133	0,000	0,109	0,121	0,000	0,010	0,000	0,000	5,792	6,000	0,447
Ítem33	0,755	-1,147	0,000	0,092	0,155	0,000	0,011	0,000	0,000	3,840	6,000	0,698
Ítem34	0,318	-3,241	0,005	0,082	1,117	0,134	0,077	0,002	0,101	2,439	6,000	0,875

ANEXO 32. *Parámetros y ajuste del modelo de 1 Parámetro de la prueba online de Secundaria*

Modelo 1 Parámetro - Online

	b	se(b)	Chi cuadrado	gl	p
Ítem1	-1,122	0,083	46,145	8,000	0,000
Ítem2	-2,405	0,111	10,149	8,000	0,255
Ítem3	-1,561	0,090	14,624	8,000	0,067
Ítem4	-1,303	0,085	19,667	8,000	0,012
Ítem5	-2,475	0,114	11,700	8,000	0,165
Ítem6	-1,152	0,083	42,469	8,000	0,000
Ítem7	-0,426	0,077	14,494	8,000	0,070
Ítem8	-0,067	0,076	11,881	8,000	0,157
Ítem9	0,764	0,078	108,103	8,000	0,000
Ítem10	-0,195	0,076	26,859	8,000	0,001
Ítem11	-1,830	0,095	16,111	8,000	0,041
Ítem12	-0,822	0,080	27,189	8,000	0,001
Ítem13	-1,830	0,095	20,976	8,000	0,007
Ítem14	-0,319	0,076	23,858	8,000	0,002
Ítem15	-0,043	0,076	9,902	8,000	0,272
Ítem16	-1,036	0,082	5,696	8,000	0,681
Ítem17	0,183	0,076	13,606	8,000	0,093
Ítem18	-1,224	0,084	24,697	8,000	0,002
Ítem19	0,899	0,080	53,560	8,000	0,000
Ítem20	-0,980	0,081	22,535	8,000	0,004
Ítem21	-1,406	0,087	58,077	8,000	0,000
Ítem22	-0,276	0,076	16,153	8,000	0,040
Ítem23	-2,070	0,101	35,549	8,000	0,000
Ítem24	-2,611	0,119	10,062	8,000	0,261
Ítem25	-0,500	0,077	24,815	8,000	0,002
Ítem26	-1,285	0,085	40,214	8,000	0,000
Ítem27	-1,452	0,088	36,116	8,000	0,000
Ítem28	-1,367	0,086	21,747	8,000	0,005
Ítem29	-1,105	0,083	10,530	8,000	0,230
Ítem30	-2,762	0,125	41,856	8,000	0,000
Ítem31	-2,286	0,107	32,099	8,000	0,000
Ítem32	-1,036	0,082	9,703	8,000	0,286
Ítem34	-0,565	0,078	10,731	8,000	0,217

ANEXO 33. *Parámetros y ajuste del modelo de 2 Parámetro de la prueba online de Secundaria*

Modelo 2 Parámetros - Online

	a	b	se(a)	se(b)	cov(a,b)	Chi cuadrado	gl	p
Ítem1	0,363	-2,934	0,083	0,662	0,052	4,102	7,000	0,768
Ítem2	1,004	-2,526	0,133	0,272	0,033	9,104	7,000	0,245
Ítem3	0,688	-2,158	0,097	0,283	0,025	17,458	7,000	0,015
Ítem4	0,998	-1,629	0,110	0,160	0,014	5,459	7,000	0,604
Ítem5	1,142	-2,420	0,144	0,237	0,030	4,202	7,000	0,756
Ítem6	0,200	-4,644	0,079	1,827	0,141	6,270	7,000	0,509
Ítem7	0,715	-0,544	0,086	0,114	0,005	25,457	7,000	0,001
Ítem8	0,667	-0,261	0,084	0,109	0,002	19,981	7,000	0,006
Ítem9	0,313	2,171	0,078	0,562	-0,041	17,805	7,000	0,013
Ítem10	0,556	-0,431	0,080	0,135	0,004	8,083	7,000	0,325
Ítem11	0,485	-3,454	0,097	0,657	0,061	4,701	7,000	0,696
Ítem12	1,201	-0,836	0,113	0,089	0,006	6,266	7,000	0,509
Ítem13	1,431	-1,618	0,143	0,124	0,014	6,407	7,000	0,493
Ítem14	1,024	-0,338	0,101	0,080	0,003	17,213	7,000	0,016
Ítem15	0,955	-0,119	0,097	0,080	0,001	10,063	7,000	0,185
Ítem16	0,639	-1,581	0,088	0,221	0,016	3,979	7,000	0,782
Ítem17	1,003	0,132	0,101	0,077	-0,001	17,313	7,000	0,015
Ítem18	1,072	-1,185	0,108	0,117	0,009	12,359	7,000	0,089
Ítem19	0,296	2,619	0,079	0,712	-0,053	12,262	7,000	0,092
Ítem20	1,333	-0,858	0,121	0,084	0,006	11,221	7,000	0,129
Ítem21	1,951	-1,008	0,174	0,074	0,007	9,907	7,000	0,194
Ítem22	0,927	-0,506	0,096	0,092	0,004	13,984	7,000	0,051
Ítem23	1,814	-1,483	0,175	0,099	0,012	5,773	7,000	0,566
Ítem24	1,212	-2,248	0,145	0,205	0,026	11,623	7,000	0,114
Ítem25	0,737	-0,792	0,088	0,126	0,007	11,811	7,000	0,107
Ítem26	1,631	-1,237	0,150	0,091	0,009	4,945	7,000	0,667
Ítem27	1,413	-1,378	0,135	0,108	0,011	12,992	7,000	0,072
Ítem28	1,104	-1,414	0,114	0,131	0,011	8,756	7,000	0,271
Ítem29	1,167	-0,938	0,112	0,096	0,007	8,399	7,000	0,299
Ítem30	2,335	-1,744	0,255	0,102	0,018	2,676	7,000	0,913
Ítem31	1,895	-1,690	0,193	0,110	0,016	7,302	7,000	0,398
Ítem32	0,918	-1,247	0,100	0,136	0,010	12,906	7,000	0,074
Ítem33	-0,266	-5,809	0,092	1,976	-0,179	17,420	7,000	0,015
Ítem34	0,687	-0,950	0,086	0,146	0,009	8,771	7,000	0,269

ANEXO 34. *Parámetros y ajuste del modelo de 3 Parámetro de la prueba online de Secundaria*

Modelo 3 Parámetros -Online

	a	b	c	se(a)	se(b)	se(c)	cov(a,b)	cov(a,c)	cov(b,c)	Chi cuadrado	gl	p
Ítem1	0,366	-2,909	0,001	0,082	0,666	0,034	0,051	0,000	0,006	3,758	6,000	0,709
Ítem2	0,981	-2,572	0,000	0,129	0,280	0,019	0,032	0,000	0,001	8,640	6,000	0,195
Ítem3	0,692	-2,145	0,000	0,095	0,277	0,001	0,024	0,000	0,000	17,283	6,000	0,008
Ítem4	1,060	-1,372	0,121	0,216	0,614	0,261	0,121	0,047	0,155	7,634	6,000	0,266
Ítem5	1,451	-1,385	0,495	0,411	0,709	0,230	0,268	0,079	0,158	4,810	6,000	0,568
Ítem6	0,199	-4,620	0,006	0,079	2,545	0,201	0,166	0,003	0,359	10,397	6,000	0,109
Ítem7	1,358	0,427	0,314	0,322	0,194	0,065	0,039	0,016	0,011	14,341	6,000	0,026
Ítem8	0,838	0,239	0,150	0,233	0,435	0,133	0,084	0,027	0,056	19,869	6,000	0,003
Ítem9	0,416	2,597	0,106	0,363	0,859	0,242	0,126	0,084	0,135	14,579	6,000	0,024
Ítem10	0,560	-0,424	0,000	0,080	0,135	0,005	0,004	0,000	0,000	8,010	6,000	0,237
Ítem11	0,476	-3,509	0,001	0,095	0,697	0,058	0,062	0,000	0,011	7,477	6,000	0,279
Ítem12	1,250	-0,730	0,049	0,198	0,260	0,116	0,044	0,018	0,028	4,817	6,000	0,568
Ítem13	1,575	-1,312	0,189	0,272	0,336	0,171	0,080	0,035	0,054	5,066	6,000	0,535
Ítem14	1,156	-0,129	0,083	0,206	0,223	0,088	0,036	0,014	0,019	12,417	6,000	0,053
Ítem15	1,129	0,123	0,094	0,196	0,188	0,071	0,025	0,011	0,012	8,375	6,000	0,212
Ítem16	0,710	-1,049	0,161	0,240	1,291	0,362	0,292	0,079	0,462	2,416	6,000	0,878
Ítem17	1,220	0,334	0,085	0,223	0,152	0,061	0,021	0,011	0,008	15,489	6,000	0,017
Ítem18	1,397	-0,567	0,263	0,281	0,306	0,119	0,074	0,026	0,034	7,192	6,000	0,303
Ítem19	0,651	2,765	0,190	0,371	0,710	0,079	-0,188	0,025	-0,016	14,428	6,000	0,025
Ítem20	1,699	-0,471	0,185	0,269	0,172	0,079	0,037	0,015	0,012	9,032	6,000	0,172
Ítem21	2,024	-0,930	0,046	0,247	0,129	0,070	0,025	0,011	0,007	5,385	6,000	0,496
Ítem22	1,162	-0,100	0,153	0,254	0,291	0,110	0,063	0,024	0,031	10,090	6,000	0,121
Ítem23	1,830	-1,422	0,051	0,261	0,238	0,149	0,053	0,027	0,032	5,654	6,000	0,463
Ítem24	1,185	-2,285	0,000	0,140	0,209	0,002	0,025	0,000	0,000	12,702	6,000	0,048
Ítem25	0,732	-0,787	0,001	0,062	NaN	NaN	-0,011	-0,006	-0,029	14,730	6,000	0,022
Ítem26	1,609	-1,236	0,000	0,143	0,081	NaN	0,008	-0,001	-0,001	4,253	6,000	0,642
Ítem27	1,446	-1,293	0,050	0,278	0,428	0,240	0,110	0,057	0,100	10,721	6,000	0,097
Ítem28	1,097	-1,413	0,001	0,127	0,221	0,087	0,022	0,005	0,015	8,712	6,000	0,190
Ítem29	1,156	-0,935	0,000	0,111	0,098	0,000	0,007	0,000	0,000	7,393	6,000	0,286
Ítem30	2,294	-1,758	0,002	0,286	0,159	0,105	0,035	0,014	0,012	5,680	6,000	0,460
Ítem31	1,857	-1,704	0,000	0,188	0,113	0,000	0,016	0,000	0,000	6,726	6,000	0,347
Ítem32	0,929	-1,170	0,034	0,541	1,972	0,807	1,055	0,429	1,587	14,056	6,000	0,029
Ítem34	0,686	-0,946	0,000	0,086	0,148	0,008	0,009	0,000	0,000	6,317	6,000	0,389

ANEXO 35. *Representación gráfica de la estructura unidimensional en la prueba de Comprensión Lectora- Primaria*

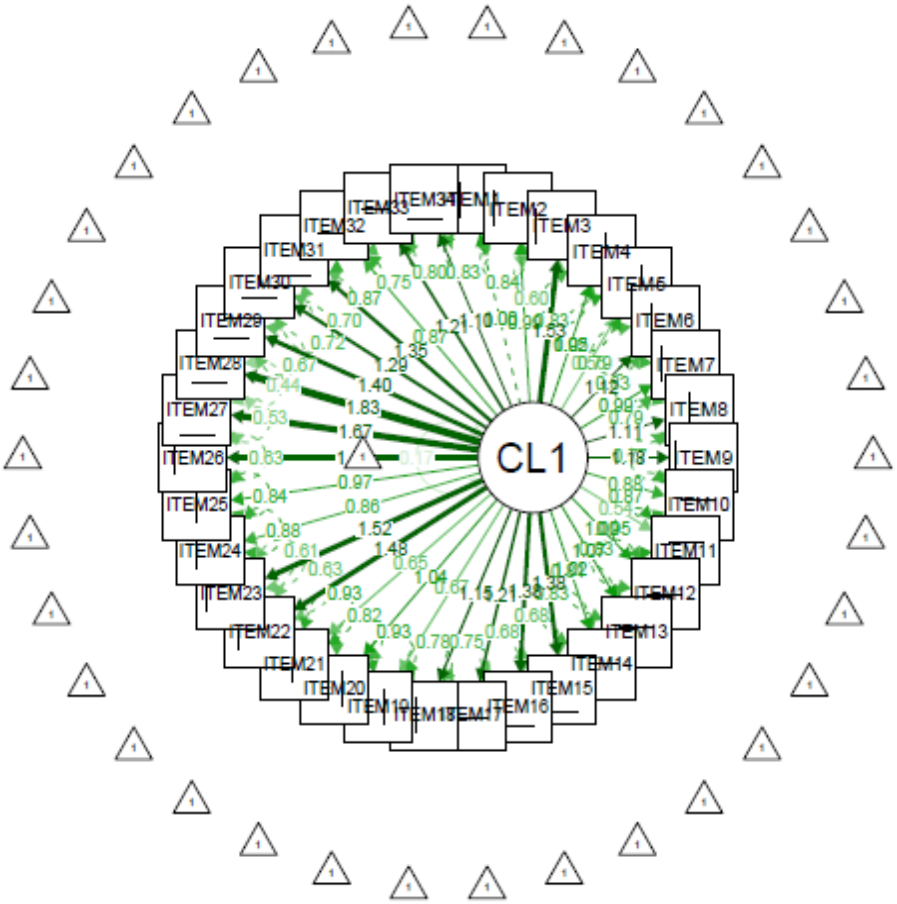


Figura 1: Estructura unidimensional prueba en papel-primaria

ANEXO 35. *Representación gráfica de la estructura unidimensional en la prueba de Comprensión Lectora- Primaria (continuación)*

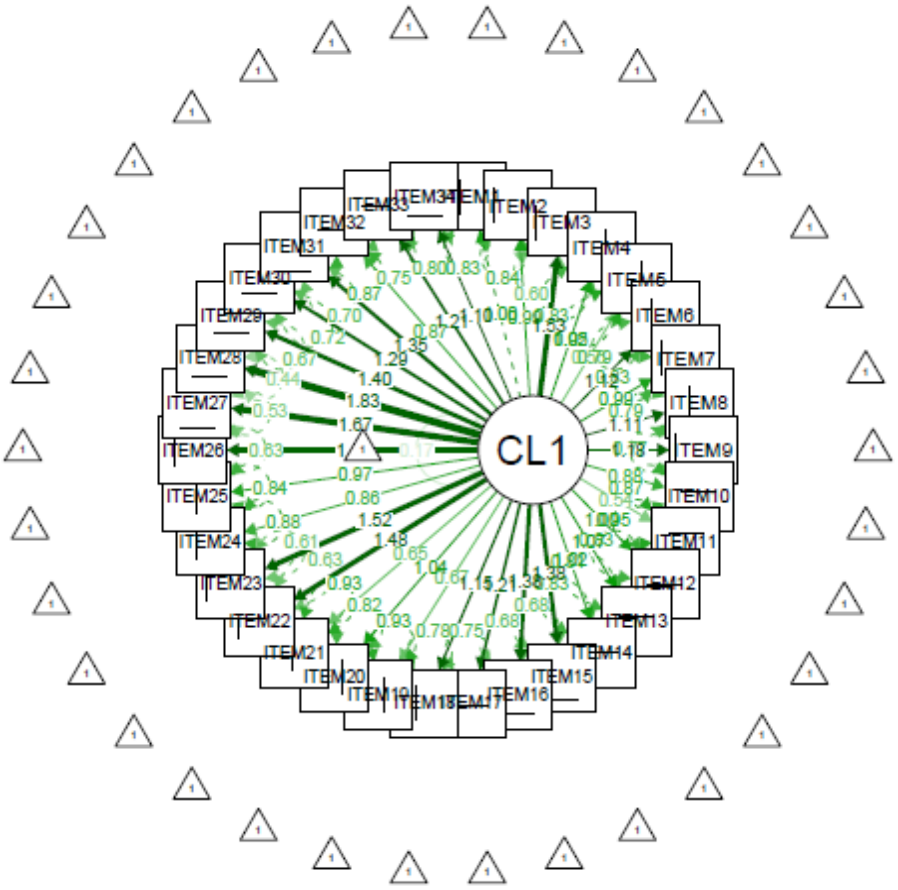


Figura 2: Estructura unidimensional prueba en online-primaria

ANEXO 36. *Representación gráfica de la estructura unidimensional en la prueba de Comprensión Lectora- Secundaria*

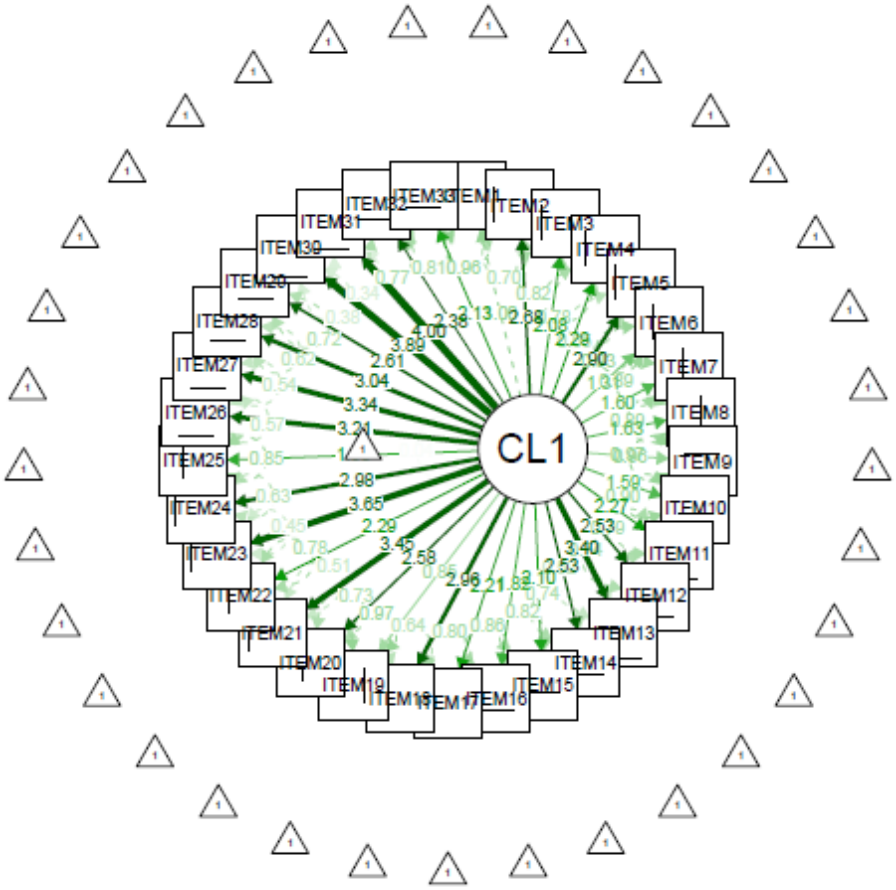


Figura 3: Estructura unidimensional prueba en papel-secundaria

ANEXO 36. *Representación gráfica de la estructura unidimensional en la prueba de Comprensión Lectora- Secundaria (continuación)*

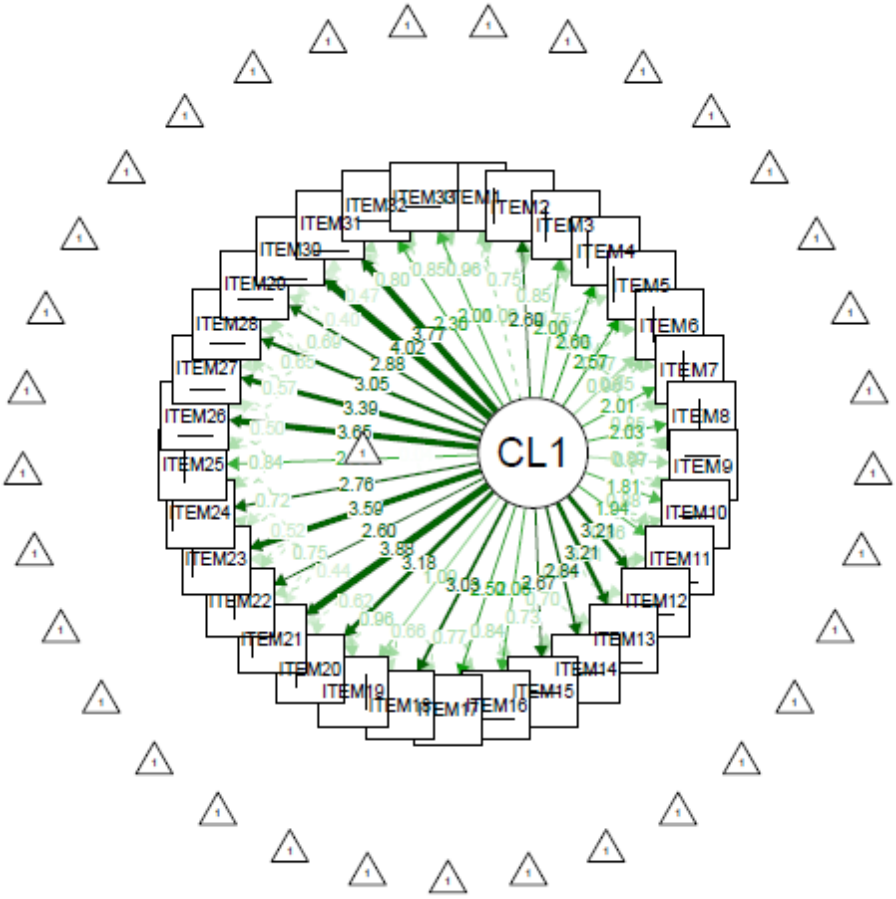


Figura 4: Estructura unidimensional prueba en online-secundaria

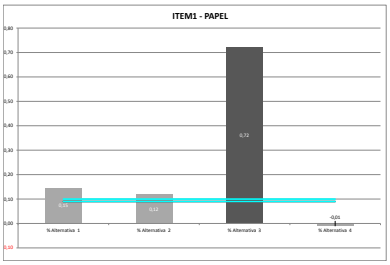
ANEXO 37. *Características y Funcionamiento Diferencial de Versiones en los ítems de la prueba de Comprensión Lectora en Secundaria.*

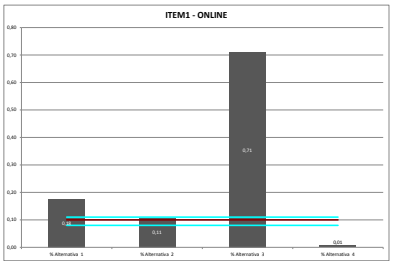
Tabla 1.
Características y Funcionamiento Diferencial de Versiones en el Ítem 1

Descripción desde la TCT

Ítem 1	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,72	0,20	0,45	0,21	0,13	3	24,5	22,1	0,72	0,451
Online	0,71	0,21	0,45	0,14	0,04	3	24,7	23,3	0,75	0,435

Porcentaje de elección de cada alternativa – ítem 1 en papel y online





Modelo TRI de 2 Parámetros

Ítem 1	Parámetro a	Parámetro b	p
Papel	0,325	-3,195	0,743
Online	0,276	-3,553	0,382

Técnicas detección DVF Ítem 1

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	DVF	DVF
0,000	0,001	0,414	Uniforme	0,0006	0,0007	0,0000	Débil					

*Diferencias significativas ($p \leq 0{,}0003$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

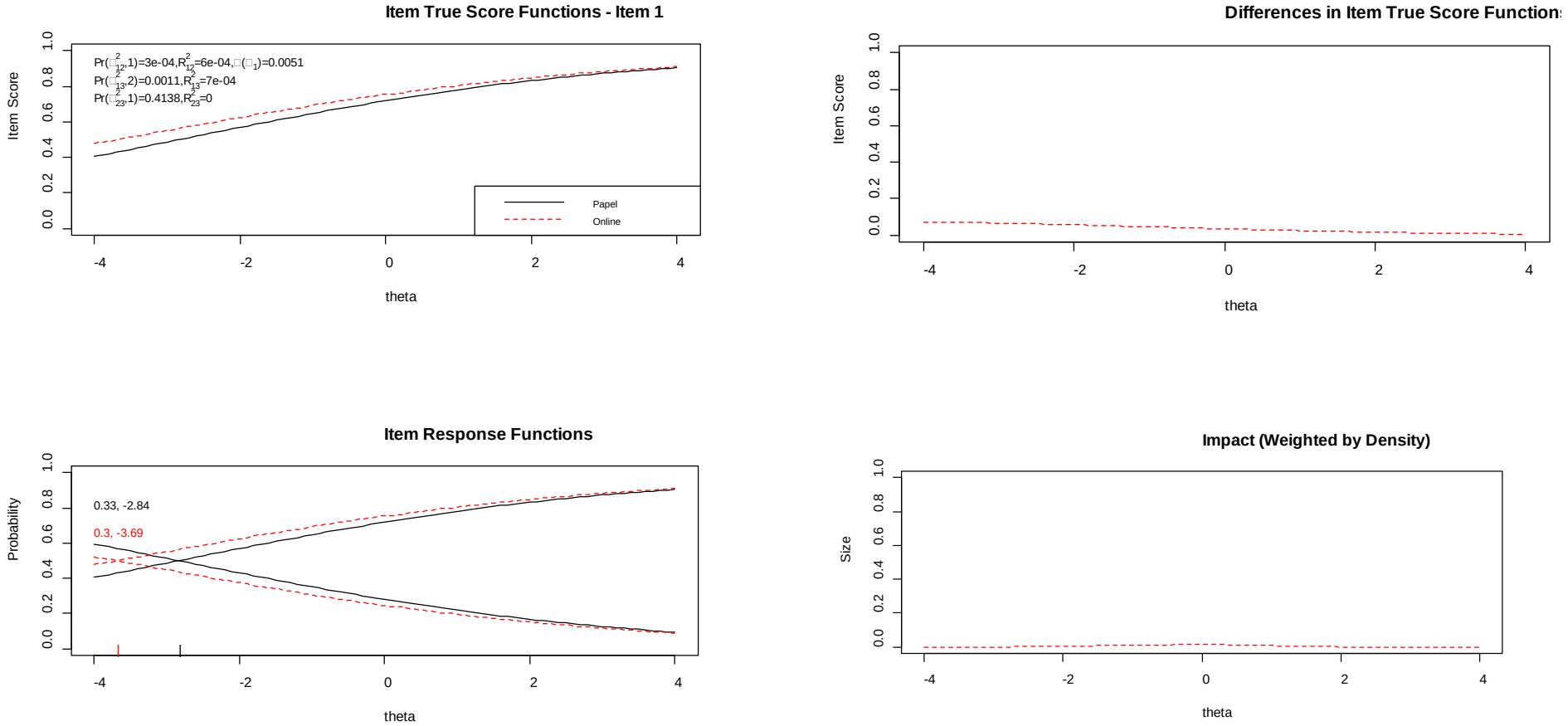


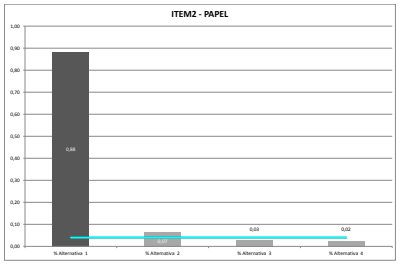
Figura 5. Características y Funcionamiento Diferencial de Versiones en el Ítem 1

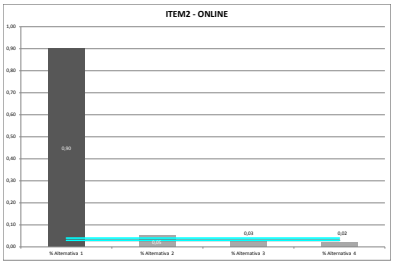
Tabla 2.
Características y Funcionamiento Diferencial de Versiones en el Ítem 2

Descripción desde la TCT

Ítem 2	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,88	0,10	0,32	0,36	0,31	1	24,5	18,7	0,88	0,325
Online	0,90	0,09	0,30	0,26	0,19	1	24,7	20,8	0,89	0,313

Porcentaje de elección de cada alternativa – ítem 2 en papel y online





Modelo TRI de 2 Parámetros

Ítem 2	Parámetro a	Parámetro b	p
Papel	1,227	-3,195	0,447
Online	1,015	-2,420	0,374

Técnicas detección DVF Ítem 2

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	DVF	DVF
0,006	0,022	0,764	Uniforme	0,0006	0,0006	0,0000	Débil					

*Diferencias significativas ($p \leq 0,0006$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

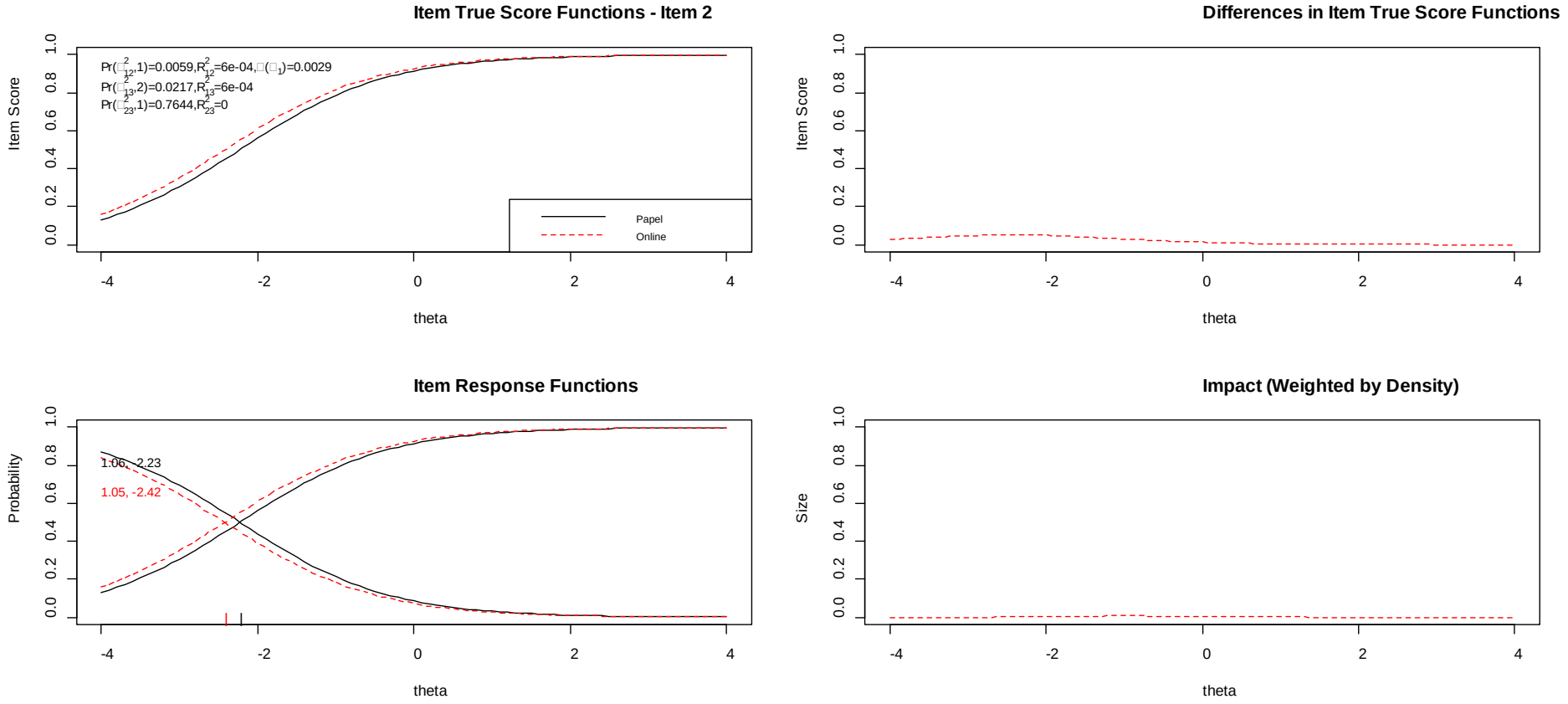


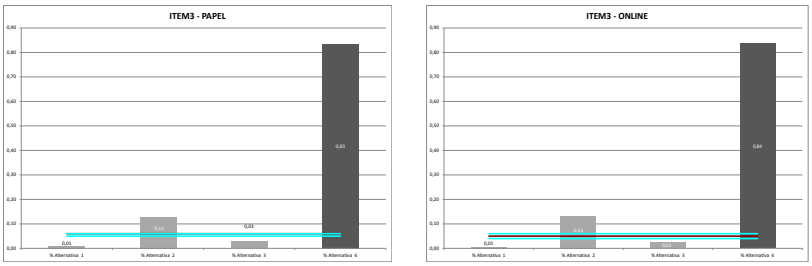
Figura 6. Características y Funcionamiento Diferencial de Versiones en el Ítem 2

Tabla 3.
Características y Funcionamiento Diferencial de Versiones en el Ítem 3

Descripción desde la TCT

Ítem 3	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,83	0,14	0,37	0,32	0,25	4	24,6	20,2	0,83	0,373
Online	0,84	0,14	0,37	0,30	0,23	4	24,9	21,2	0,80	0,402

Porcentaje de elección de cada alternativa – ítem 3 en papel y online



Modelo TRI de 2 Parámetros

Ítem 3	Parámetro a	Parámetro b	p
Papel	0,939	-2,342	0,242
Online	0,681	-2,156	0,003

Técnicas detección DVF Ítem 3

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	DVF
0,001	0,003	0,293	Uniforme	0,0007	0,0007	0,0001	Débil					

*Diferencias significativas ($p \leq 0{,}0009$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

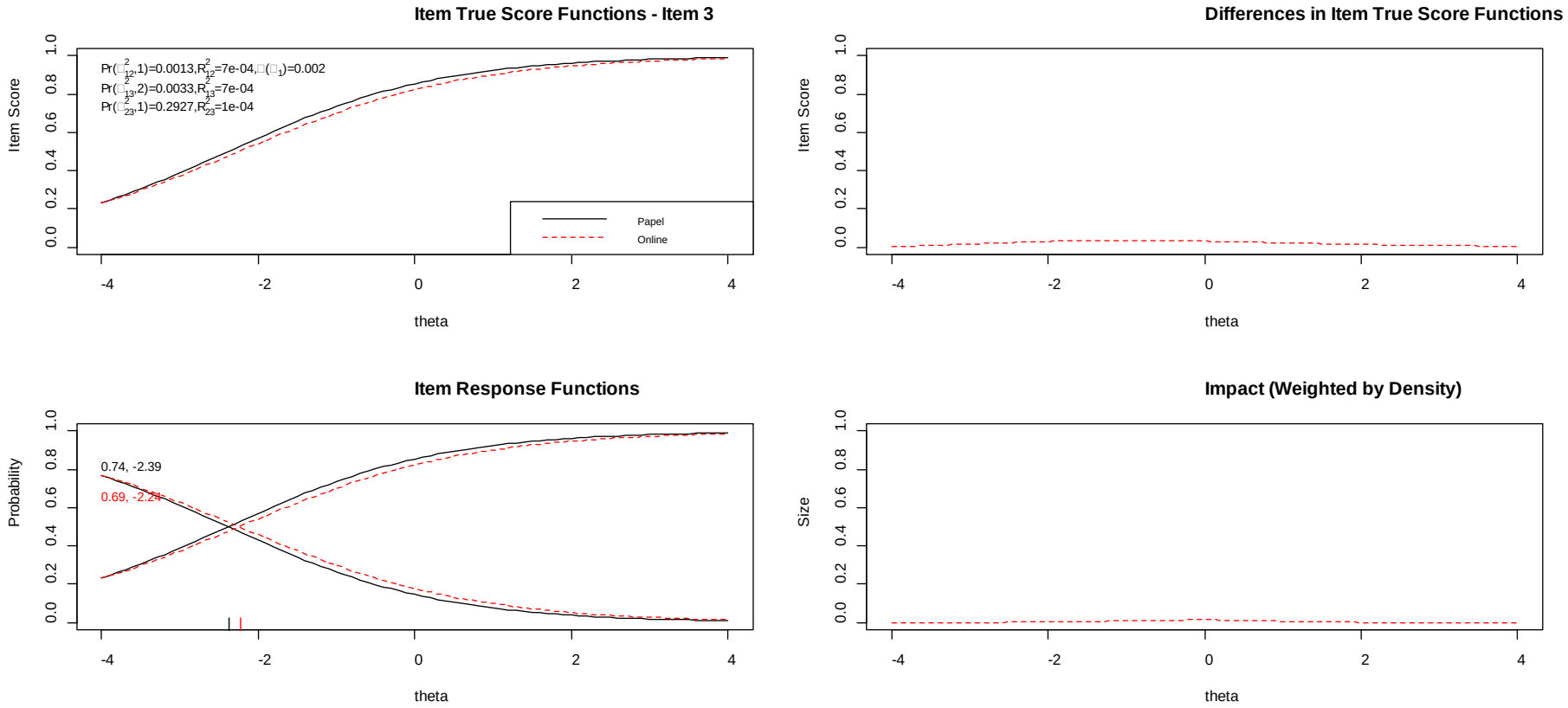


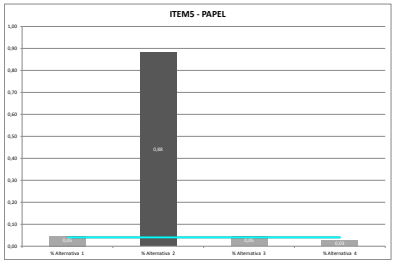
Figura 7. Características y Funcionamiento Diferencial de Versiones en el Ítem 3

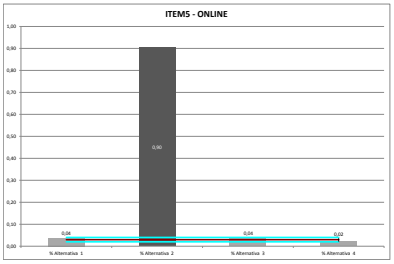
Tabla 4.
Características y Funcionamiento Diferencial de Versiones en el Ítem 5

Descripción desde la TCT

Ítem 5	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,88	0,10	0,32	0,39	0,34	2	24,6	18,3	0,88	0,327
Online	0,90	0,09	0,29	0,34	0,28	2	24,8	19,5	0,90	0,297

Porcentaje de elección de cada alternativa – ítem 5 en papel y online





Modelo TRI de 2 Parámetros

Ítem 5	Parámetro a	Parámetro b	p
Papel	1,512	-1,827	0,803
Online	1,256	-2,152	0,734

Técnicas detección DVF Ítem 5

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	DVF	DVF
0,000	0,000	0,050	Uniforme	0,0022	0,0025	0,0003	Débil					

*Diferencias significativas ($p \leq 0{,}0015$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

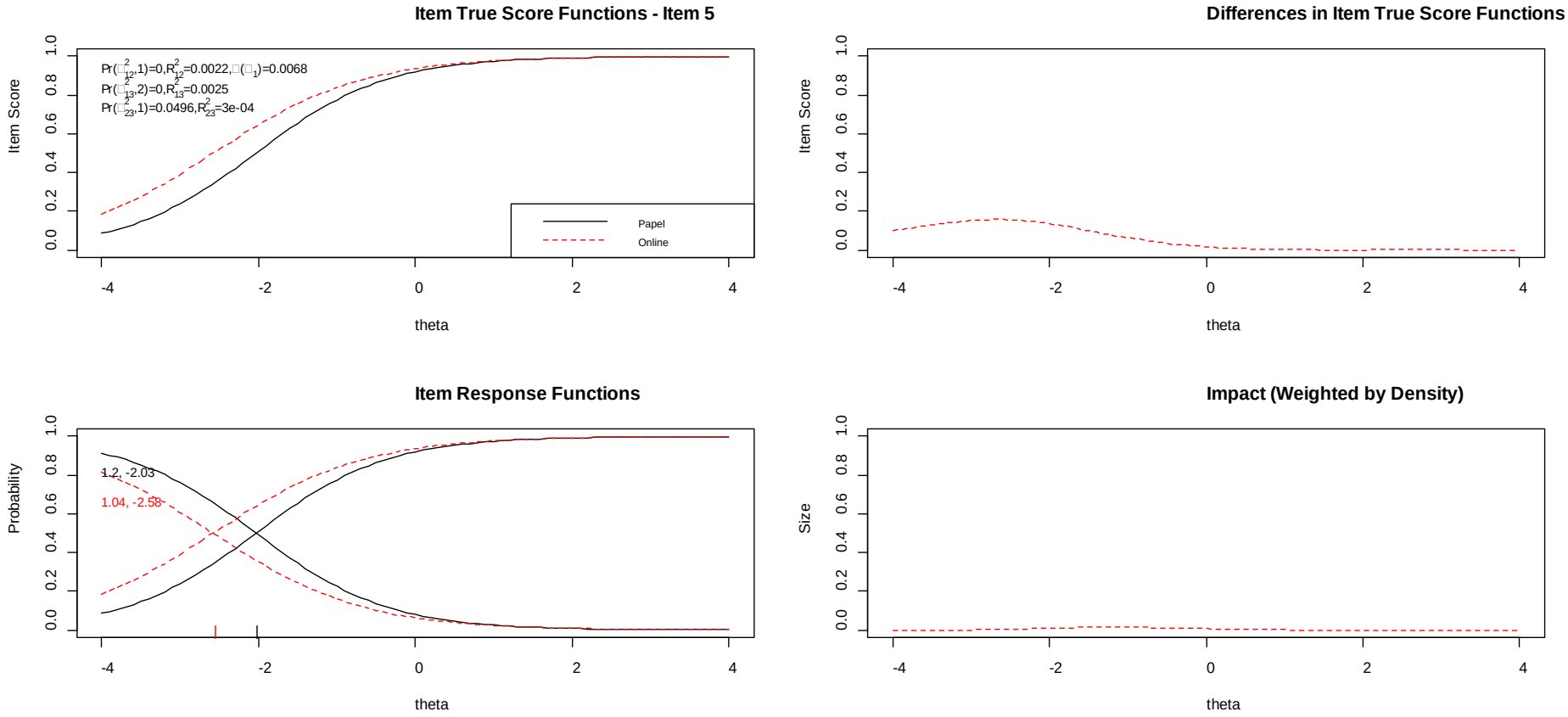


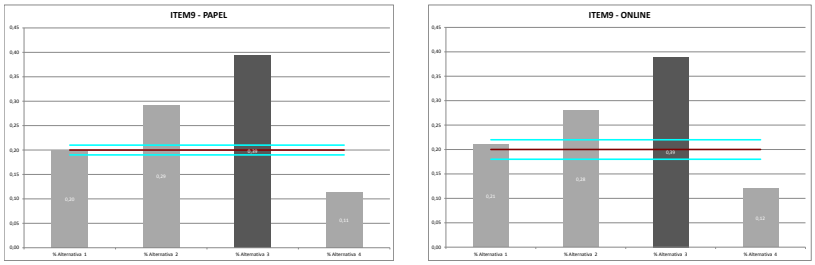
Figura 8. Características y Funcionamiento Diferencial de Versiones en el Ítem 5

Tabla 5.
Características y Funcionamiento Diferencial de Versiones en el Ítem 9

Descripción desde la TCT

Ítem 9	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,39	0,24	0,49	0,22	0,13	3	25,2	22,9	0,40	0,491
Online	0,39	0,24	0,49	0,22	0,12	3	25,6	23,5	0,35	0,478

Porcentaje de elección de cada alternativa – ítem 9 en papel y online



Modelo TRI de 2 Parámetros

Ítem 9	Parámetro a	Parámetro b	p
Papel	0,350	0,854	0,166
Online	0,085	7,576	0,063

Técnicas detección DVF Ítem 9

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	DVF
0,000	0,000	0,261	Uniforme	0,0007	0,0008	0,0001	Débil					

*Diferencias significativas ($p \leq 0{,}0027$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

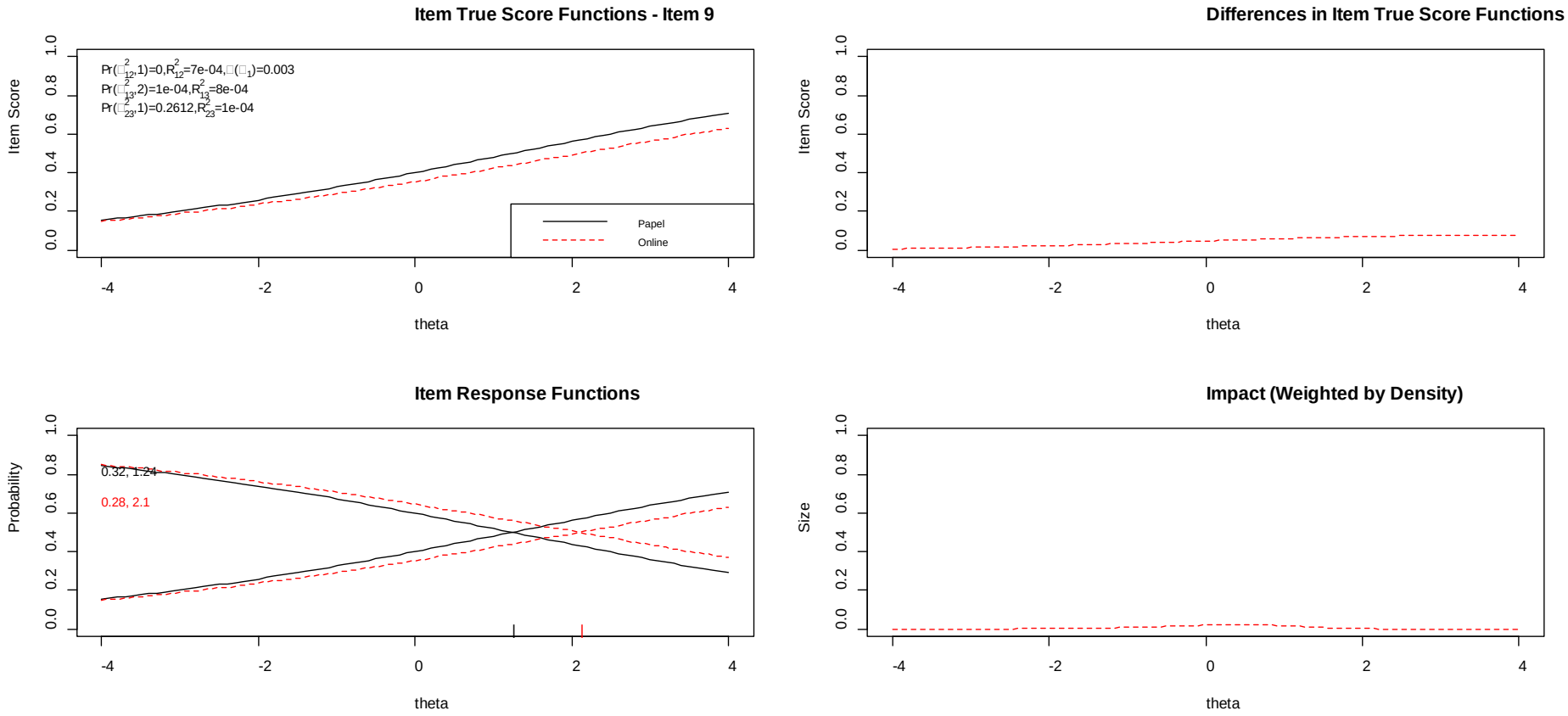


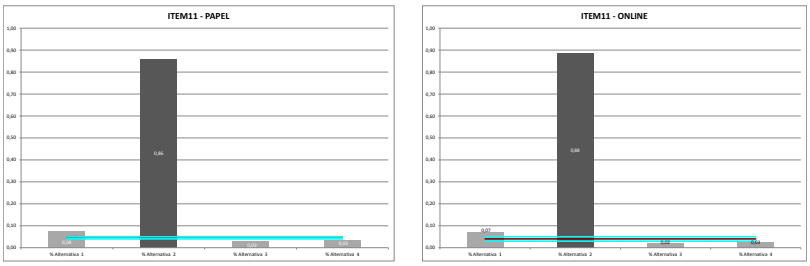
Figura 9. Características y Funcionamiento Diferencial de Versiones en el Ítem 9

Tabla 6.
Características y Funcionamiento Diferencial de Versiones en el Ítem 11

Descripción desde la TCT

Ítem 11	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,86	0,12	0,35	0,36	0,30	2	24,6	19,3	0,86	0,342
Online	0,88	0,10	0,32	0,33	0,26	2	24,9	20,2	0,83	0,376

Porcentaje de elección de cada alternativa – ítem 11 en papel y online



Modelo TRI de 2 Parámetros

Ítem 11	Parámetro a	Parámetro b	p
Papel	0,774	-2,827	0,445
Online	0,602	-2,808	0,057

Técnicas detección DVF Ítem 11

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	DVF
0,001	0,000	0,001	No Uniforme	0,0009	0,0016	0,0008	Débil					

*Diferencias significativas ($p \leq 0,0033$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

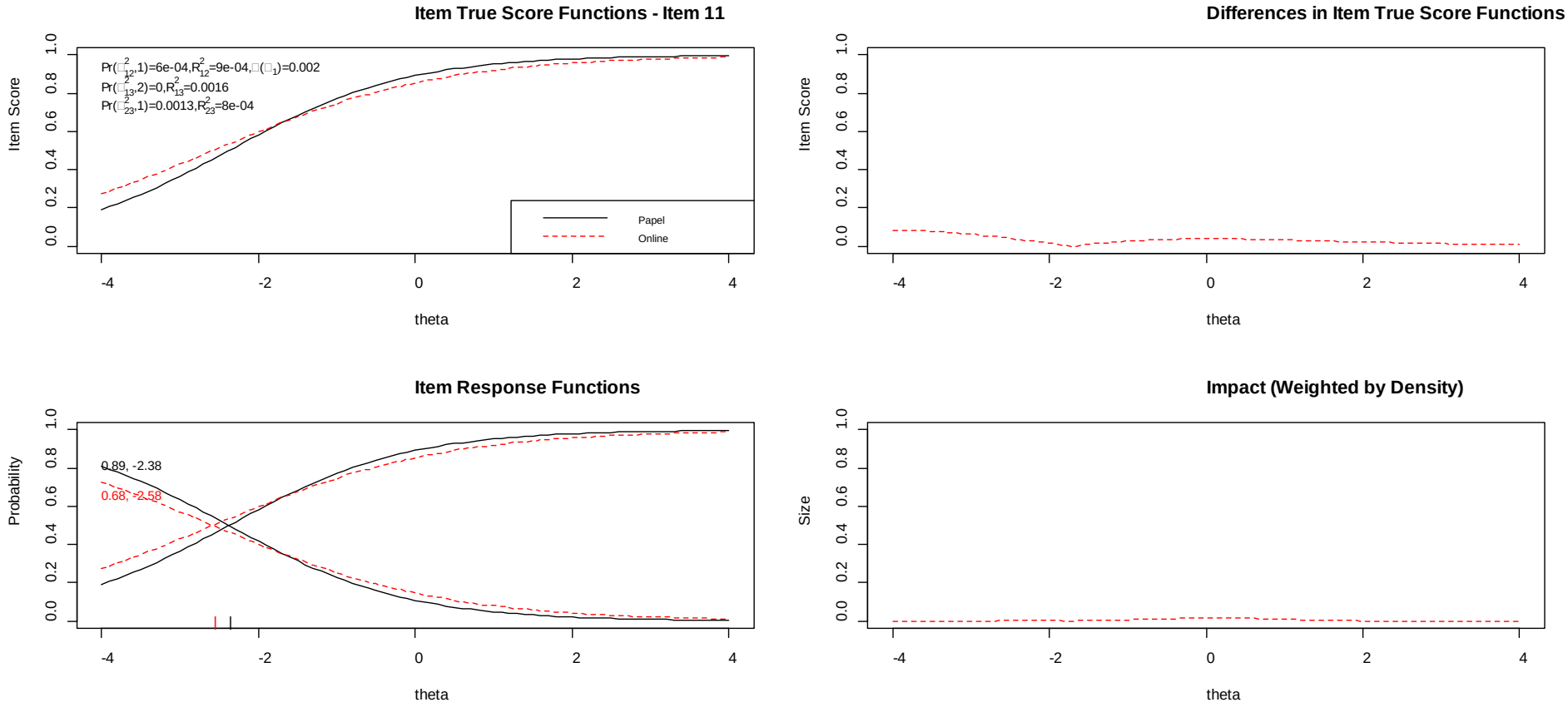


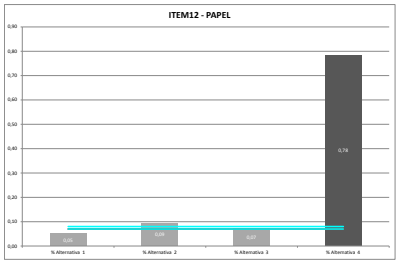
Figura 10. Características y Funcionamiento Diferencial de Versiones en el Ítem 11

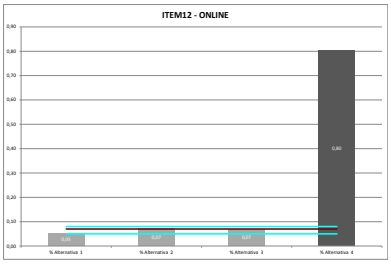
Tabla 7.
Características y Funcionamiento Diferencial de Versiones en el Ítem 12

Descripción desde la TCT

Ítem 12	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,78	0,17	0,41	0,40	0,33	4	24,9	19,9	0,79	0,405
Online	0,80	0,16	0,40	0,40	0,32	4	25,2	20,6	0,68	0,468

Porcentaje de elección de cada alternativa – ítem 12 en papel y online





Modelo TRI de 2 Parámetros

Ítem 12	Parámetro a	Parámetro b	p
Papel	1,155	-1,508	0,653
Online	1,239	-0,733	0,120

Técnicas detección DVF Ítem 12

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,000	No Uniforme	0,0074	0,0086	0,0012	Débil					

*Diferencias significativas ($p \leq 0,0036$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

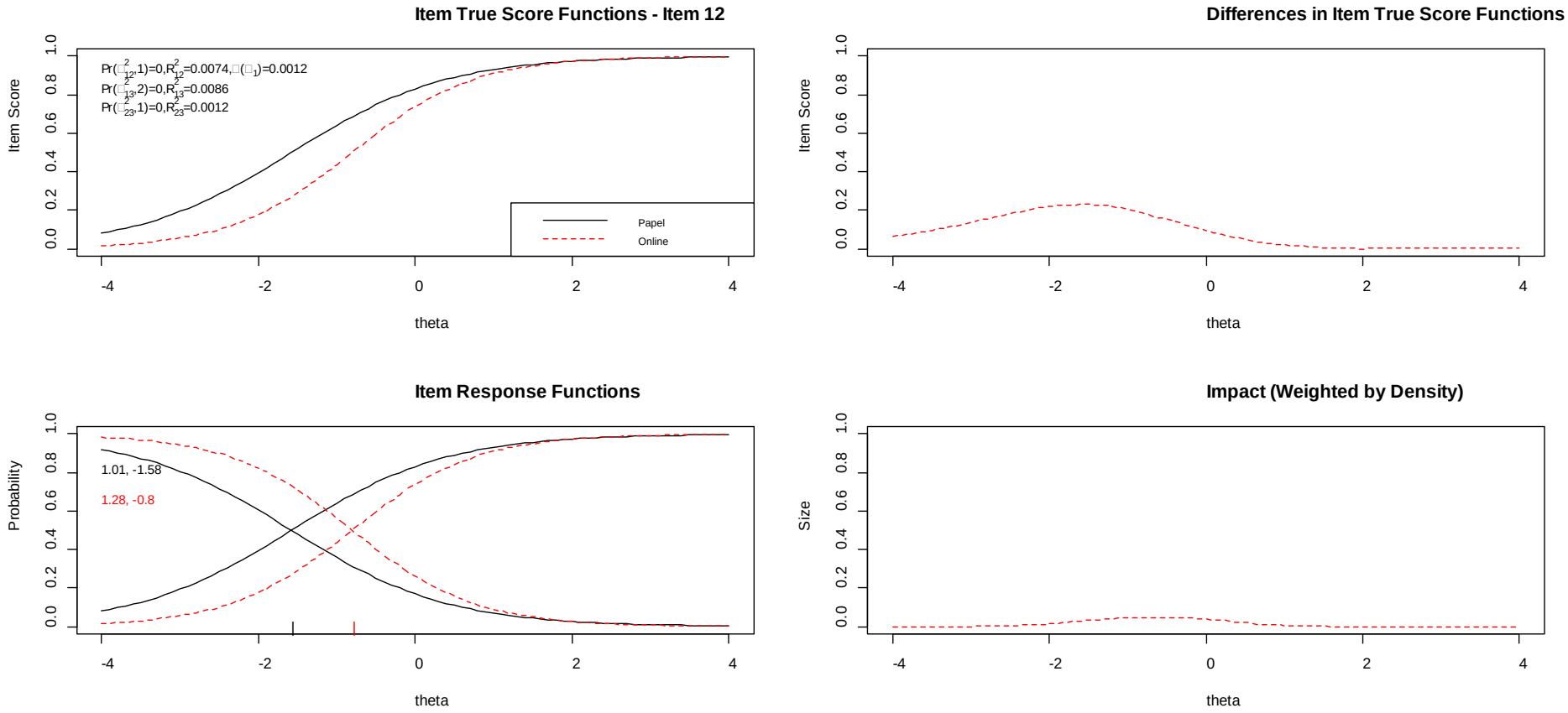


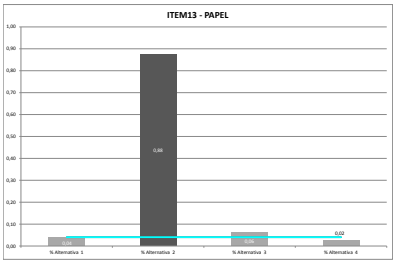
Figura 11. Características y Funcionamiento Diferencial de Versiones en el Ítem 12

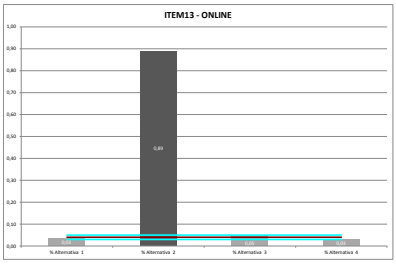
Tabla 8.
Características y Funcionamiento Diferencial de Versiones en el Ítem 13

Descripción desde la TCT

Ítem 13	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,88	0,11	0,33	0,48	0,42	2	24,8	17,3	0,87	0,331
Online	0,89	0,10	0,31	0,41	0,35	2	25,0	19,0	0,84	0,367

Porcentaje de elección de cada alternativa – ítem 13 en papel y online





Modelo TRI de 2 Parámetros

Ítem 13	Parámetro a	Parámetro b	p
Papel	1,853	-1,603	0,535
Online	1,364	-1,524	0,735

Técnicas detección DVF Ítem 13

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	No DVF
0,003	0,000	0,009	No Uniforme	0,0007	0,0012	0,0005	Débil					

*Diferencias significativas ($p \leq 0,0039$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

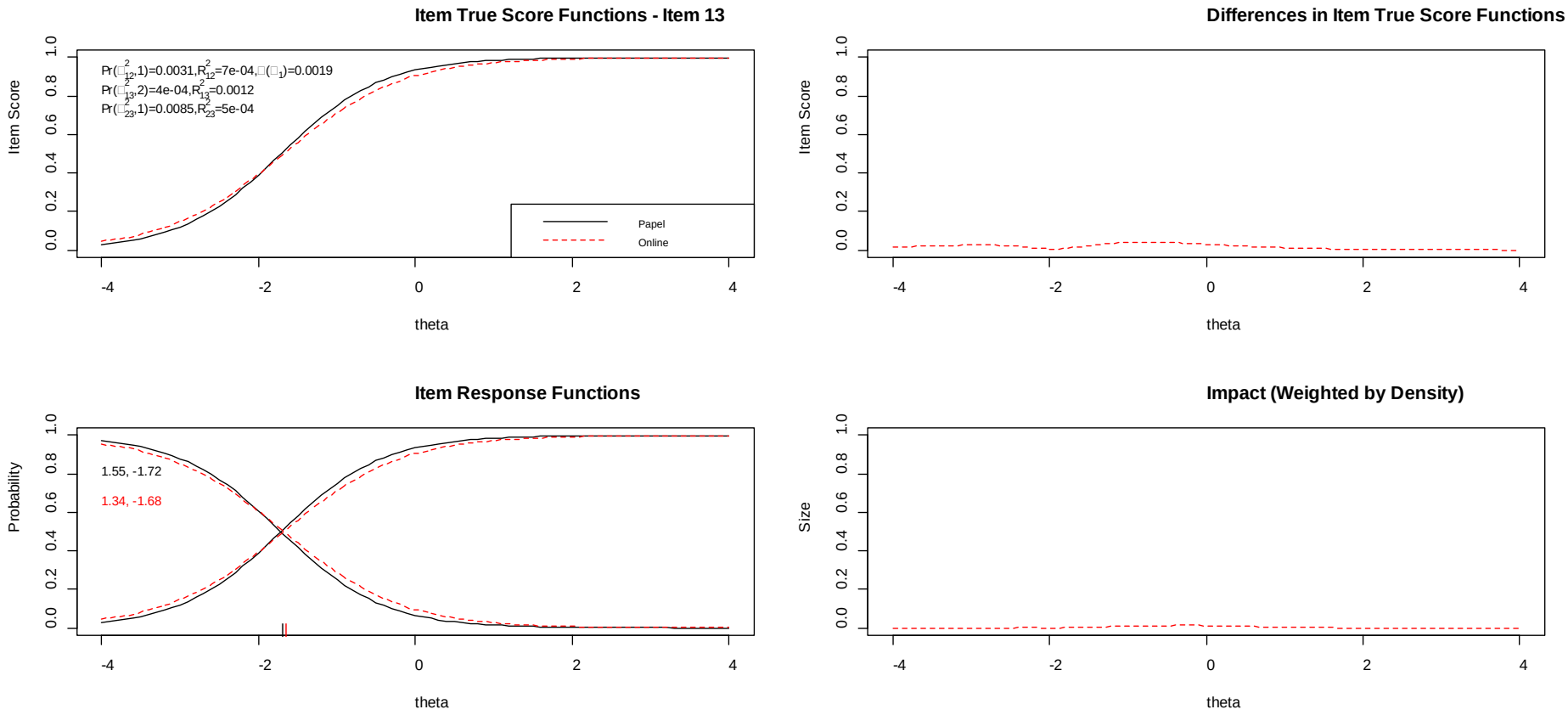


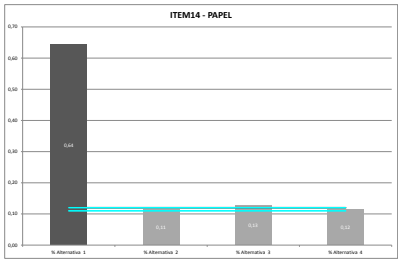
Figura 12. Características y Funcionamiento Diferencial de Versiones en el Ítem 13

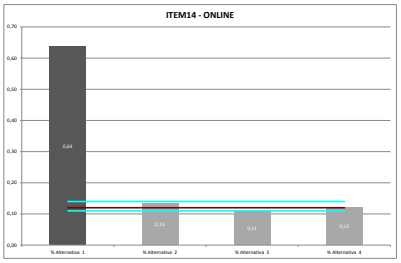
Tabla 9.
Características y Funcionamiento Diferencial de Versiones en el Ítem 14

Descripción desde la TCT

Ítem 14	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,64	0,23	0,48	0,45	0,37	1	25,6	20,7	0,66	0,475
Online	0,64	0,23	0,48	0,43	0,34	1	25,8	21,7	0,56	0,497

Porcentaje de elección de cada alternativa – ítem 14 en papel y online





Modelo TRI de 2 Parámetros

Ítem 14	Parámetro a	Parámetro b	p
Papel	1,229	-0,702	0,155
Online	1,204	-0,292	0,044

Técnicas detección DVF Ítem 14

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,736	Uniforme	0,0030	0,0031	0,0000	Débil					

*Diferencias significativas ($p \leq 0,0042$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

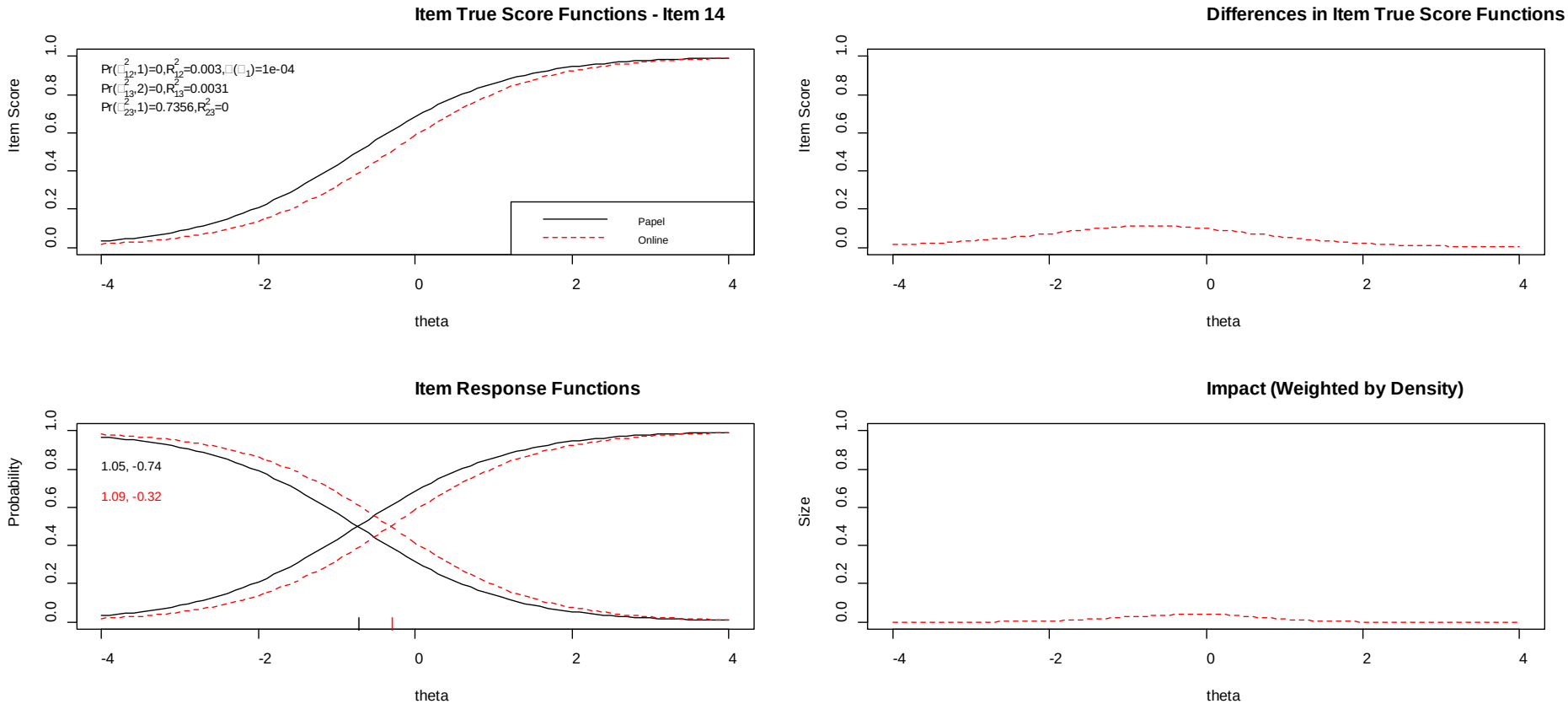


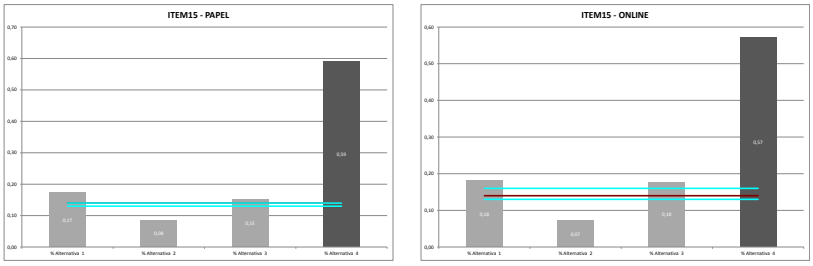
Figura 13. Características y Funcionamiento Diferencial de Versiones en el Ítem 14

Tabla 10.
Características y Funcionamiento Diferencial de Versiones en el Ítem 15

Descripción desde la TCT

Ítem 15	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,59	0,24	0,49	0,41	0,32	4	25,6	21,3	0,61	0,488
Online	0,57	0,24	0,49	0,36	0,26	4	25,8	22,4	0,51	0,500

Porcentaje de elección de cada alternativa – ítem 15 en papel y online



Modelo TRI de 2 Parámetros

Ítem 15	Parámetro a	Parámetro b	p
Papel	0,805	-0,873	0,122
Online	0,908	-0,055	0,125

Técnicas detección DVF Ítem 15

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,000	No Uniforme	0,0027	0,0034	0,0007	Débil					

*Diferencias significativas ($p \leq 0,0045$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

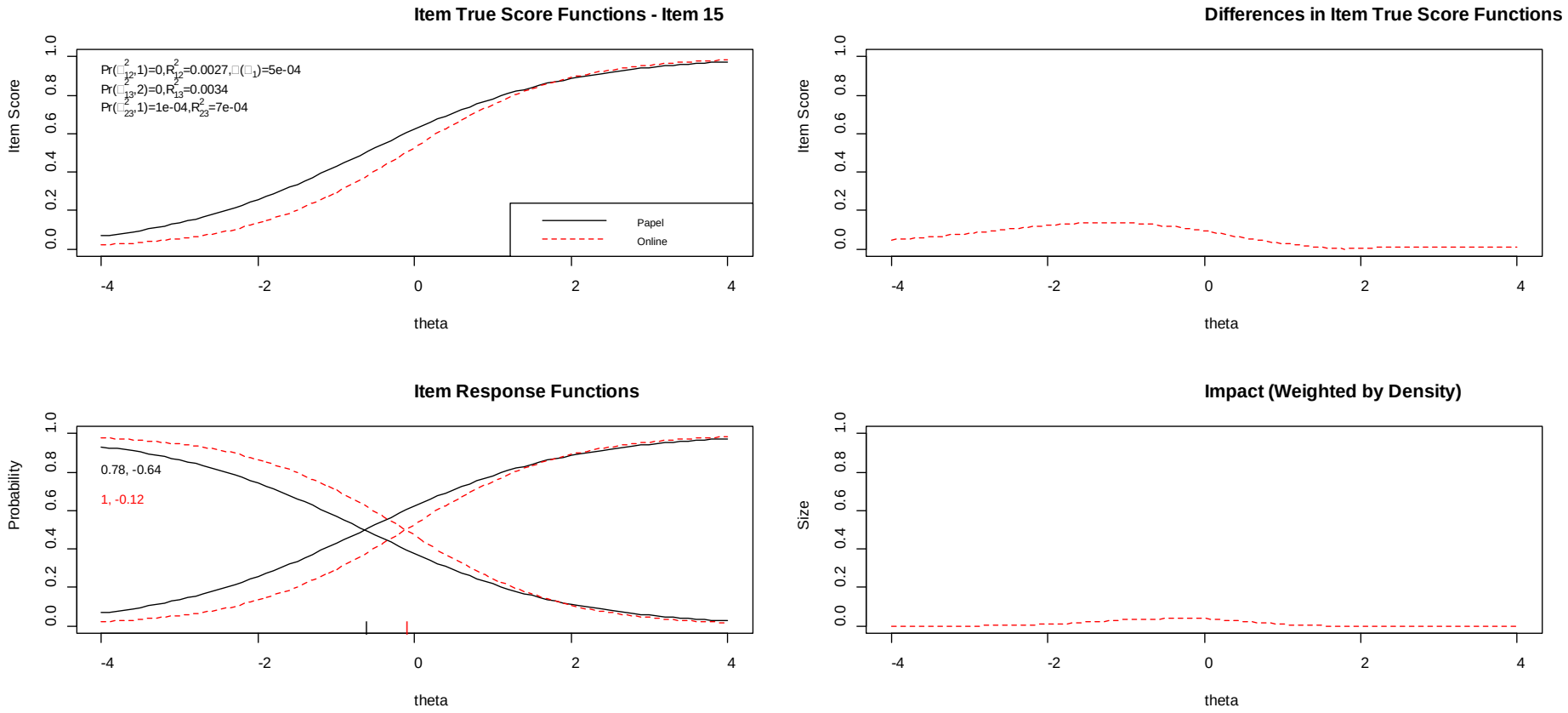


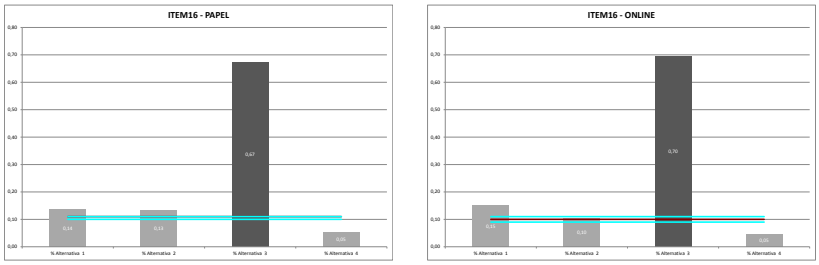
Figura 14. Características y Funcionamiento Diferencial de Versiones en el Ítem 15

Tabla 11.
Características y Funcionamiento Diferencial de Versiones en el Ítem 16

Descripción desde la TCT

Ítem 16	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,67	0,22	0,47	0,33	0,24	3	25,0	21,4	0,67	0,470
Online	0,70	0,21	0,46	0,27	0,18	3	25,2	22,4	0,71	0,453

Porcentaje de elección de cada alternativa – ítem 16 en papel y online



Modelo TRI de 2 Parámetros

Ítem 16	Parámetro a	Parámetro b	p
Papel	0,658	-1,218	0,450
Online	0,717	-1,375	0,031

Técnicas detección DVF Ítem 16

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,618	Uniforme	0,0013	0,0013	0,0000	Débil					

*Diferencias significativas ($p \leq 0{,}0048$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

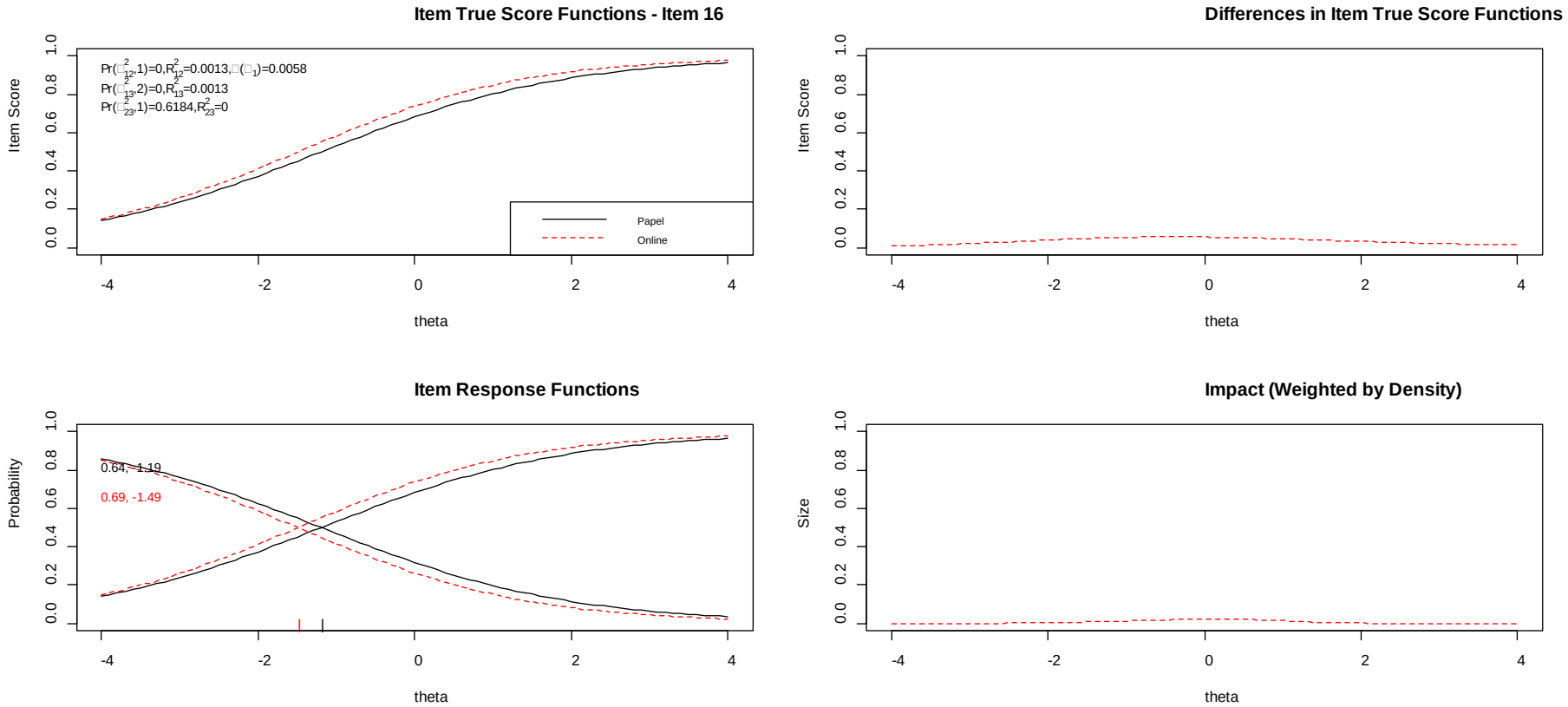


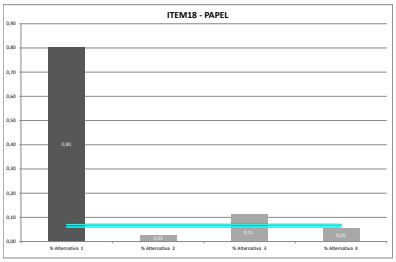
Figura 15. Características y Funcionamiento Diferencial de Versiones en el Ítem 16

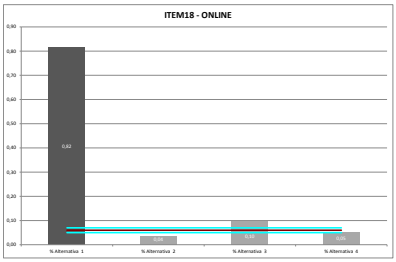
Tabla 12.
Características y Funcionamiento Diferencial de Versiones en el Ítem 18

Descripción desde la TCT

Ítem 18	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,80	0,16	0,40	0,46	0,40	1	25,0	19,0	0,81	0,394
Online	0,82	0,15	0,39	0,43	0,35	1	25,3	20,2	0,74	0,438

Porcentaje de elección de cada alternativa – ítem 18 en papel y online





Modelo TRI de 2 Parámetros

Ítem 18	Parámetro a	Parámetro b	p
Papel	1,399	-1,481	0,129
Online	1,283	-1,063	0,177

Técnicas detección DVF Ítem 18

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,201	Uniforme	0,0021	0,0022	0,0001	Débil					

*Diferencias significativas ($p \leq 0{,}0055$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

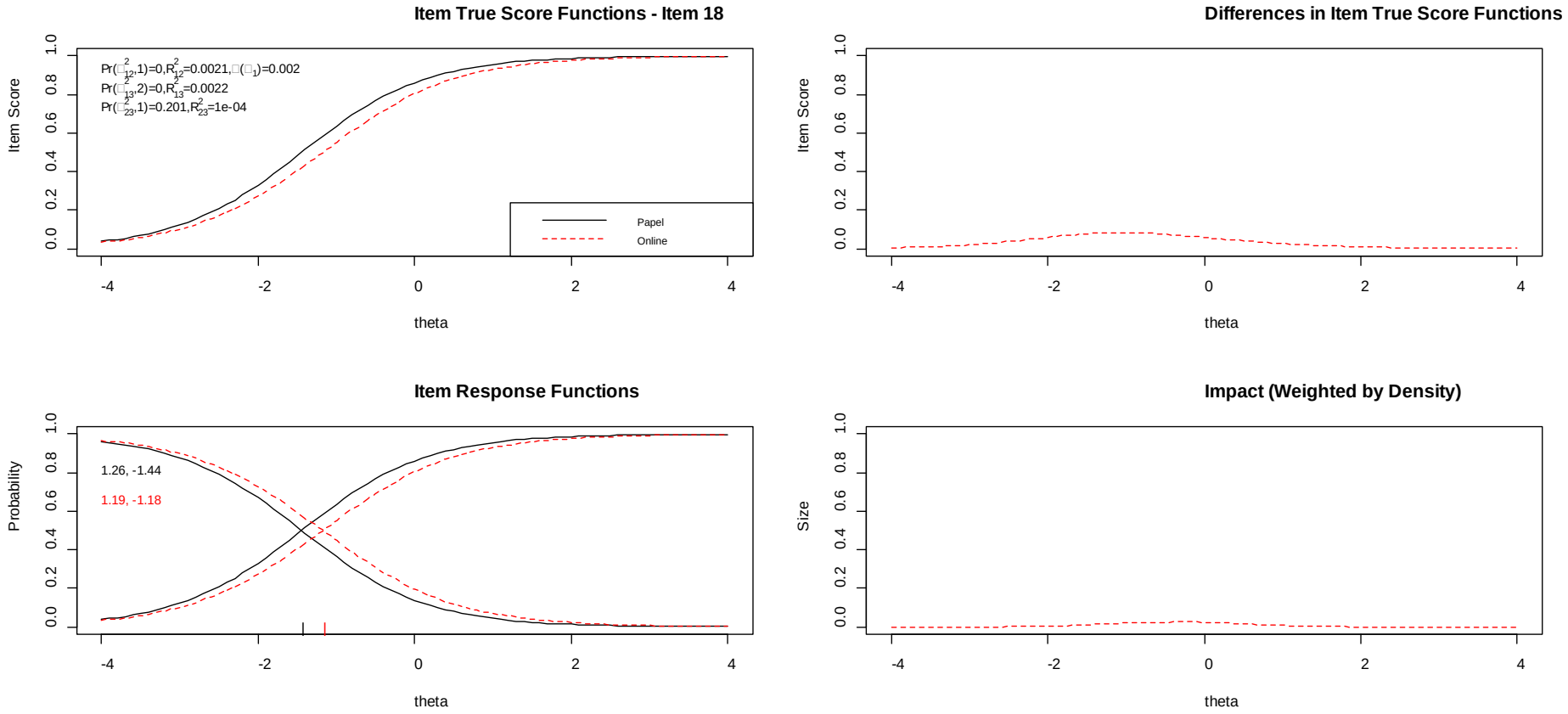


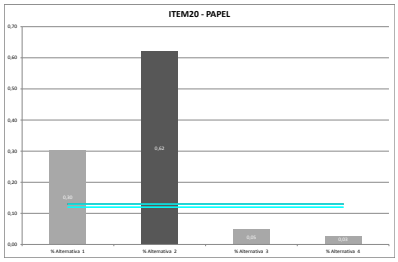
Figura 16. Características y Funcionamiento Diferencial de Versiones en el Ítem 18

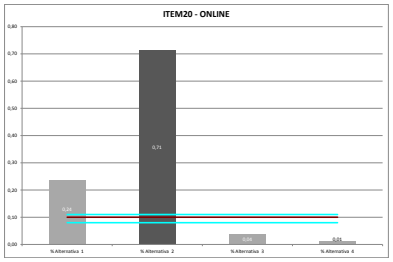
Tabla 13.
Características y Funcionamiento Diferencial de Versiones en el Ítem 20

Descripción desde la TCT

Ítem 20	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,62	0,24	0,48	0,45	0,37	2	25,7	20,8	0,62	0,485
Online	0,71	0,20	0,45	0,41	0,32	2	25,5	21,4	0,71	0,456

Porcentaje de elección de cada alternativa – ítem 20 en papel y online





Modelo TRI de 2 Parámetros

Ítem 20	Parámetro a	Parámetro b	p
Papel	1,284	-0,691	0,089
Online	1,203	-0,888	0,112

Técnicas detección DVF Ítem 20

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	DVF	DVF	DVF	DVF
0,000	0,000	0,121	Uniforme	0,0056	0,0057	0,0001	Débil					

*Diferencias significativas ($p \leq 0,0061$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

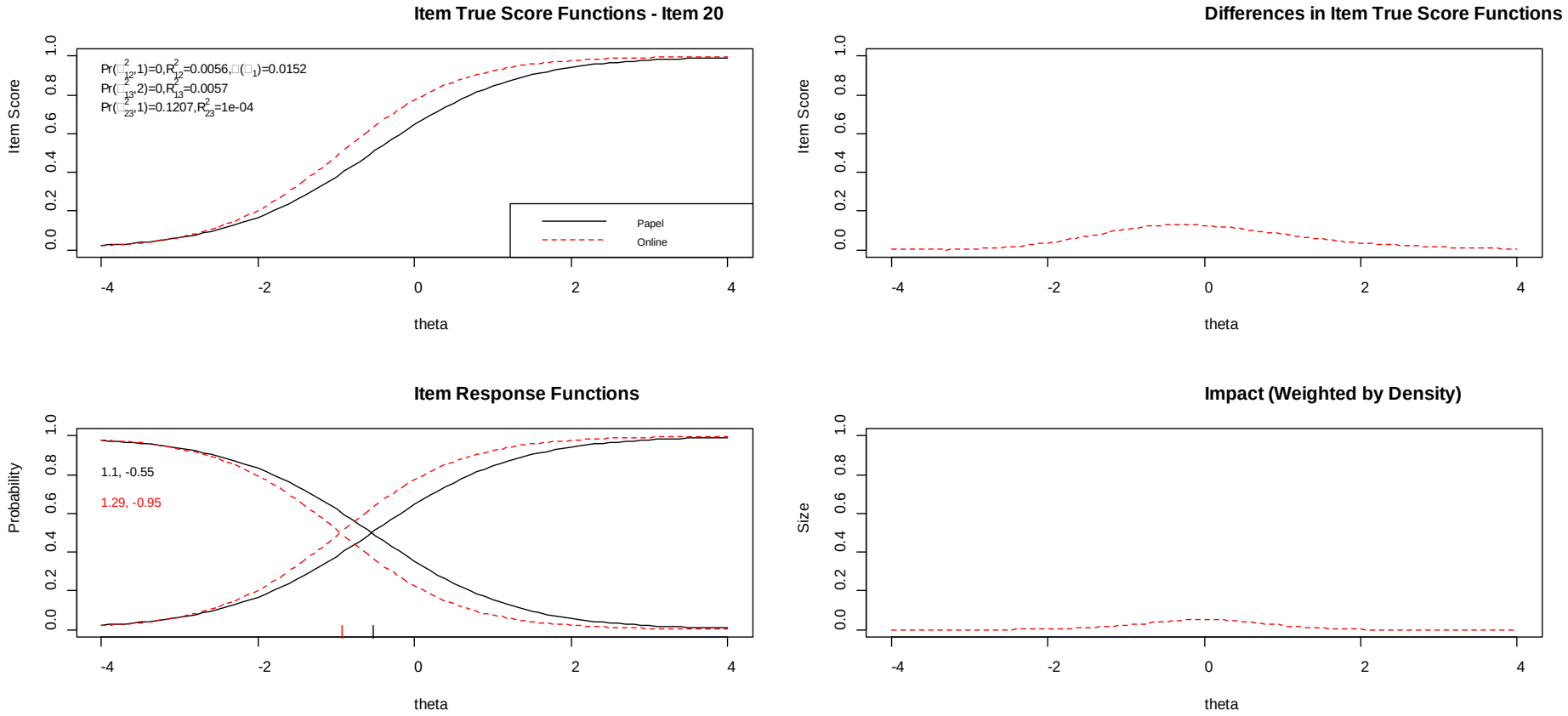


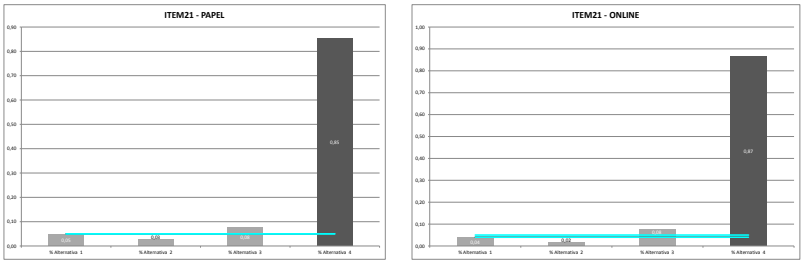
Figura 17. Características y Funcionamiento Diferencial de Versiones en el Ítem 20

Tabla 14.
Características y Funcionamiento Diferencial de Versiones en el Ítem 21

Descripción desde la TCT

Ítem 21	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,85	0,13	0,36	0,49	0,43	4	24,9	17,9	0,85	0,357
Online	0,87	0,12	0,34	0,48	0,42	4	25,2	18,7	0,77	0,419

Porcentaje de elección de cada alternativa – ítem 21 en papel y online



Modelo TRI de 2 Parámetros

Ítem 21	Parámetro a	Parámetro b	p
Papel	1,617	-1,642	0,831
Online	1,939	-0,984	0,293

Técnicas detección DVF Ítem 21

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,001	No Uniforme	0,0042	0,005	0,0007	Débil					

*Diferencias significativas ($p \leq 0,0064$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

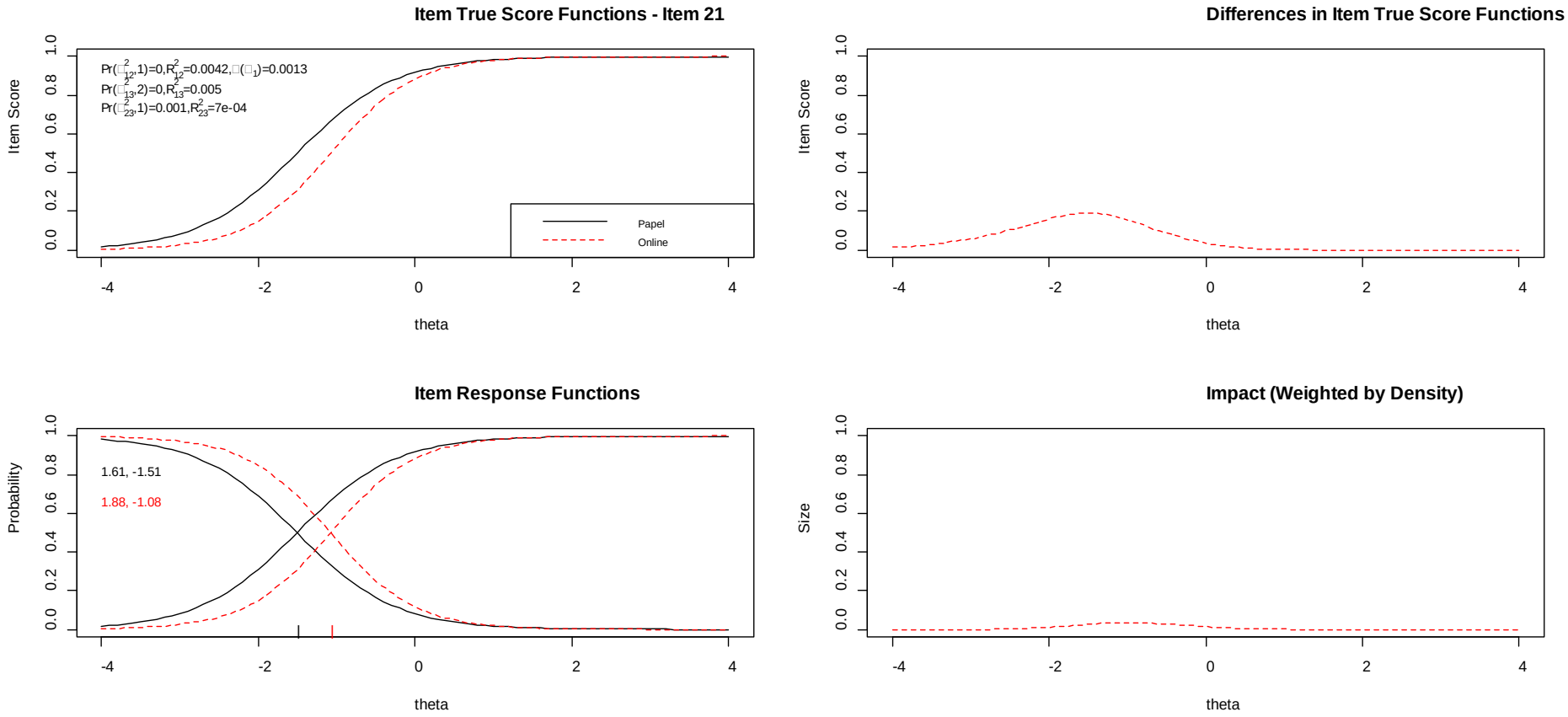


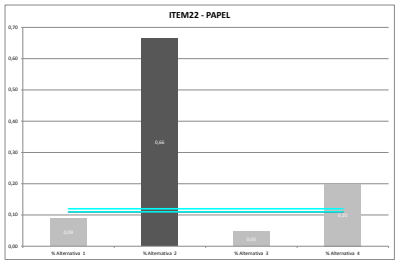
Figura 18. Características y Funcionamiento Diferencial de Versiones en el Ítem 21

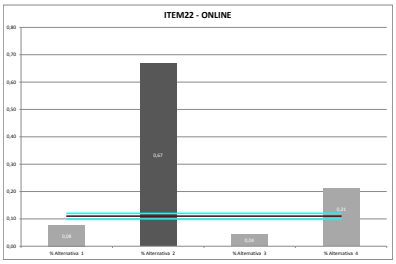
Tabla 15.
Características y Funcionamiento Diferencial de Versiones en el Ítem 22

Descripción desde la TCT

Ítem 22	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,66	0,22	0,47	0,37	0,29	2	25,2	21,1	0,66	0,474
Online	0,67	0,22	0,47	0,38	0,29	2	25,6	21,9	0,59	0,493

Porcentaje de elección de cada alternativa – ítem 22 en papel y online





Modelo TRI de 2 Parámetros

Ítem 22	Parámetro a	Parámetro b	p
Papel	0,910	-1,133	0,002
Online	0,883	-0,317	0,080

Técnicas detección DVF Ítem 22

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,373	Uniforme	0,0015	0,0016	0,0000	Débil					

*Diferencias significativas ($p \leq 0,0067$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

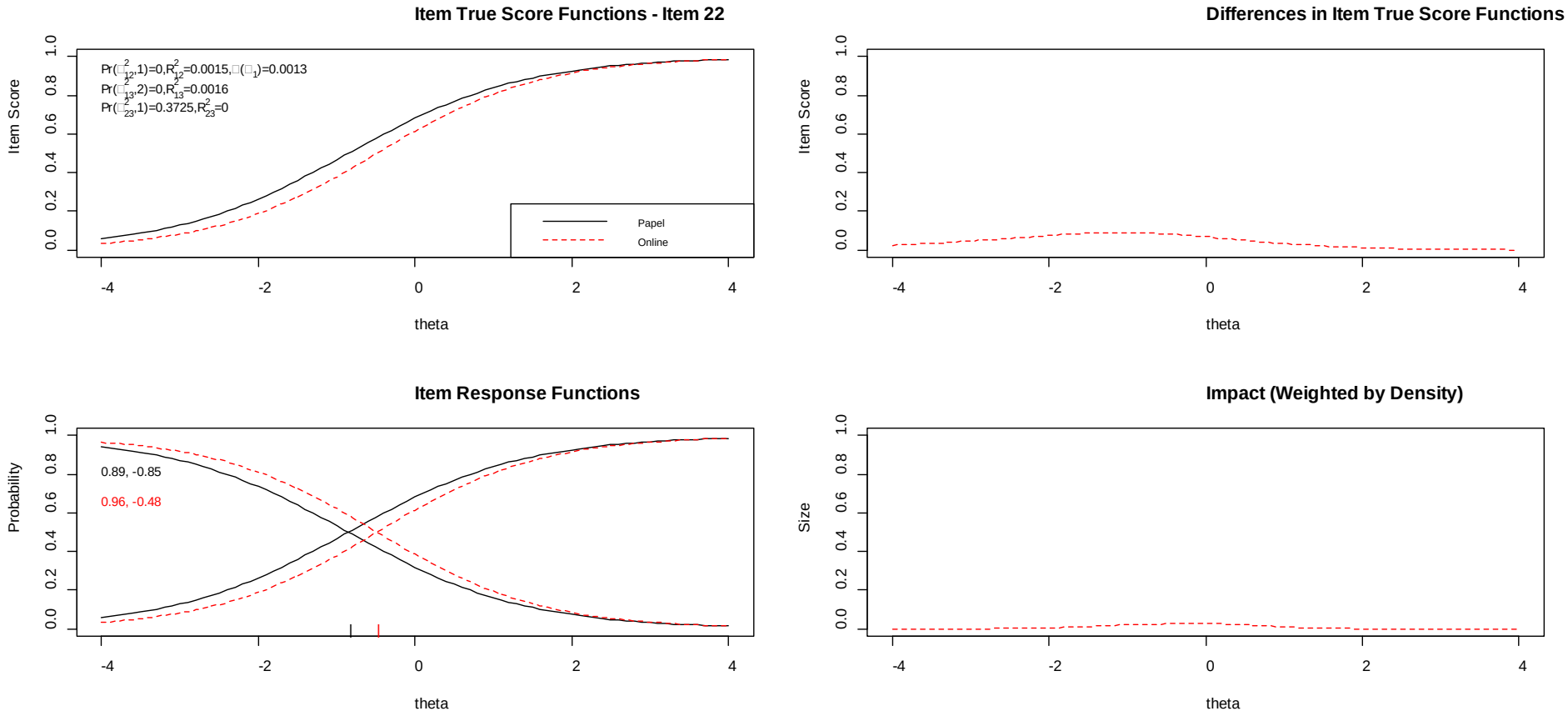


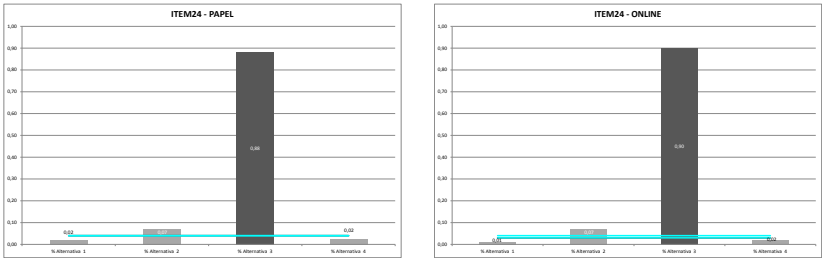
Figura 19. Características y Funcionamiento Diferencial de Versiones en el Ítem 22

Tabla 16.
Características y Funcionamiento Diferencial de Versiones en el Ítem 24

Descripción desde la TCT

Ítem 24	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,88	0,10	0,32	0,36	0,31	3	24,5	18,7	0,87	0,331
Online	0,90	0,09	0,30	0,25	0,19	3	24,7	20,8	0,91	0,292

Porcentaje de elección de cada alternativa – ítem 24 en papel y online



Modelo TRI de 2 Parámetros

Ítem 24	Parámetro a	Parámetro b	p
Papel	1,265	-2,064	0,000
Online	1,081	-2,508	0,010

Técnicas detección DVF Ítem 24

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,417	Uniforme	0,0030	0,0030	0,0001	Débil					

*Diferencias significativas ($p \leq 0,0073$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

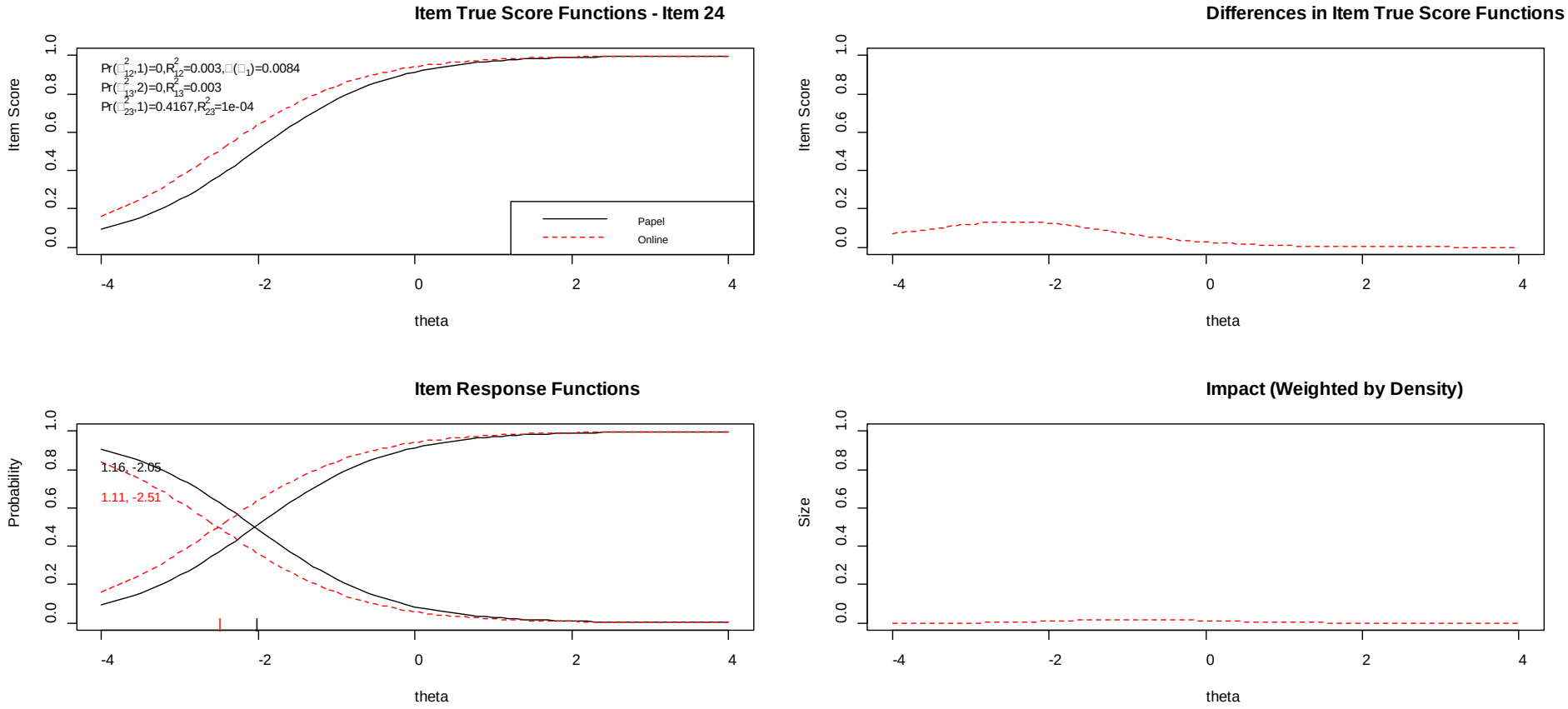


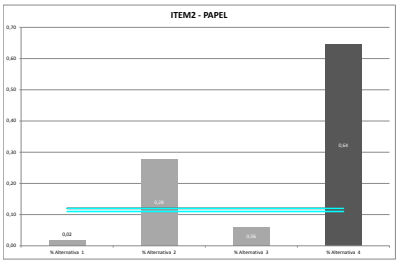
Figura 20. Características y Funcionamiento Diferencial de Versiones en el Ítem 24

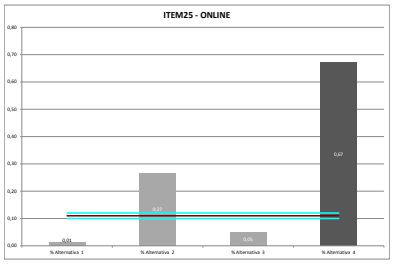
Tabla 17.
Características y Funcionamiento Diferencial de Versiones en el Ítem 25

Descripción desde la TCT

Ítem 25	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,64	0,23	0,48	0,35	0,27	4	25,2	21,4	0,66	0,474
Online	0,67	0,22	0,47	0,35	0,26	4	25,5	22,0	0,61	0,487

Porcentaje de elección de cada alternativa – ítem 25 en papel y online





Modelo TRI de 2 Parámetros

Ítem 25	Parámetro a	Parámetro b	p
Papel	0,709	-1,179	0,155
Online	0,650	-0,727	0,020

Técnicas detección DVF Ítem 25

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	No DVF
0,001	0,003	0,299	Uniforme	0,0005	0,0005	0,0001	Débil					

*Diferencias significativas ($p \leq 0,0076$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

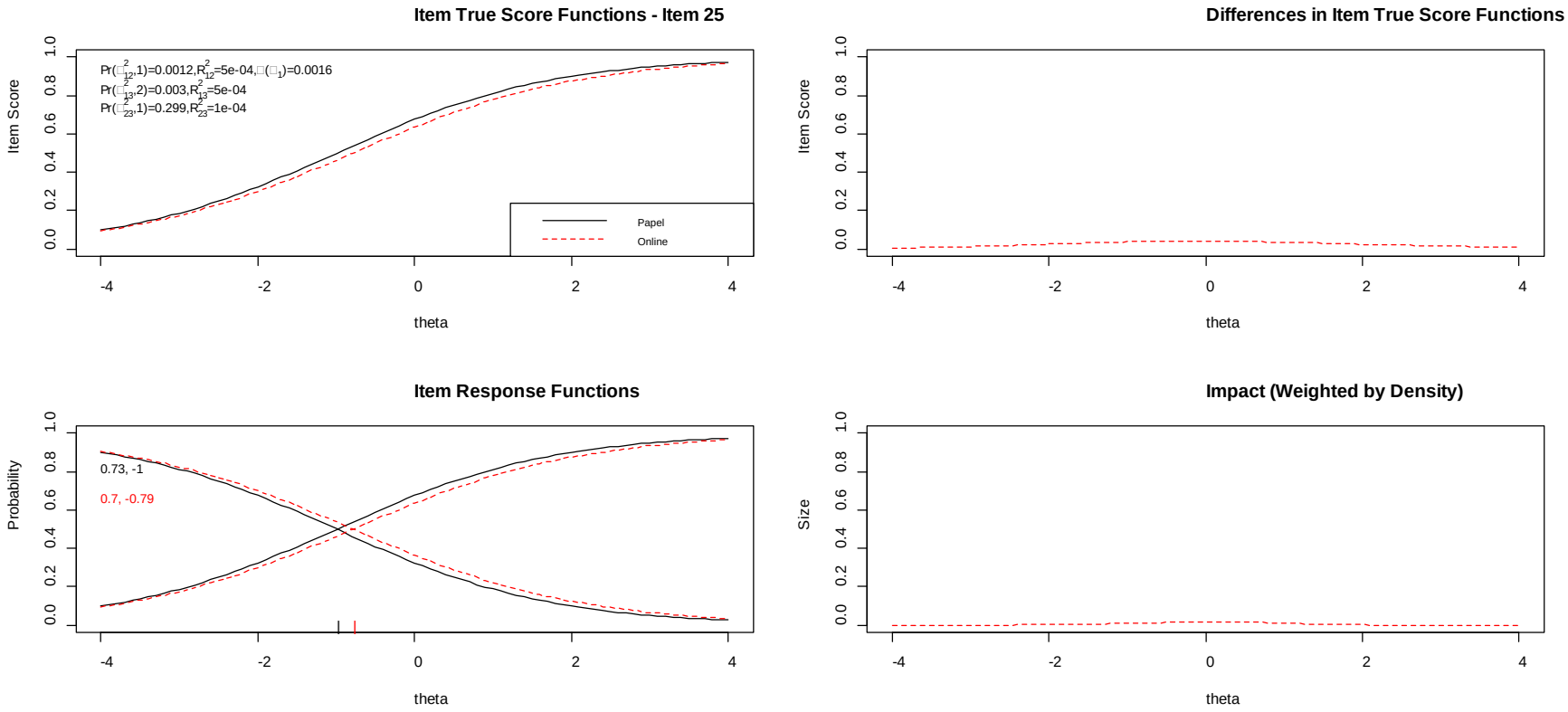


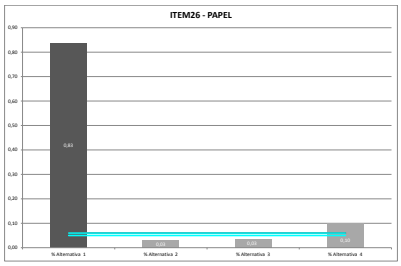
Figura 21. Características y Funcionamiento Diferencial de Versiones en el Ítem 25

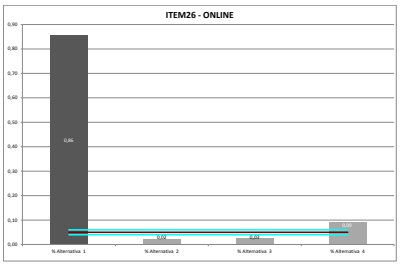
Tabla 18.
Características y Funcionamiento Diferencial de Versiones en el Ítem 26

Descripción desde la TCT

Ítem 26	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,83	0,14	0,37	0,48	0,42	1	24,9	18,3	0,83	0,372
Online	0,86	0,12	0,35	0,44	0,38	1	25,2	19,4	0,78	0,415

Porcentaje de elección de cada alternativa – ítem 26 en papel y online





Modelo TRI de 2 Parámetros

Ítem 26	Parámetro a	Parámetro b	p
Papel	1,403	-1,622	0,684
Online	1,590	-0,988	0,153

Técnicas detección DVF Ítem 26

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,015	Uniforme	0,0016	0,002	0,0004	Débil					

*Diferencias significativas ($p \leq 0,0079$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

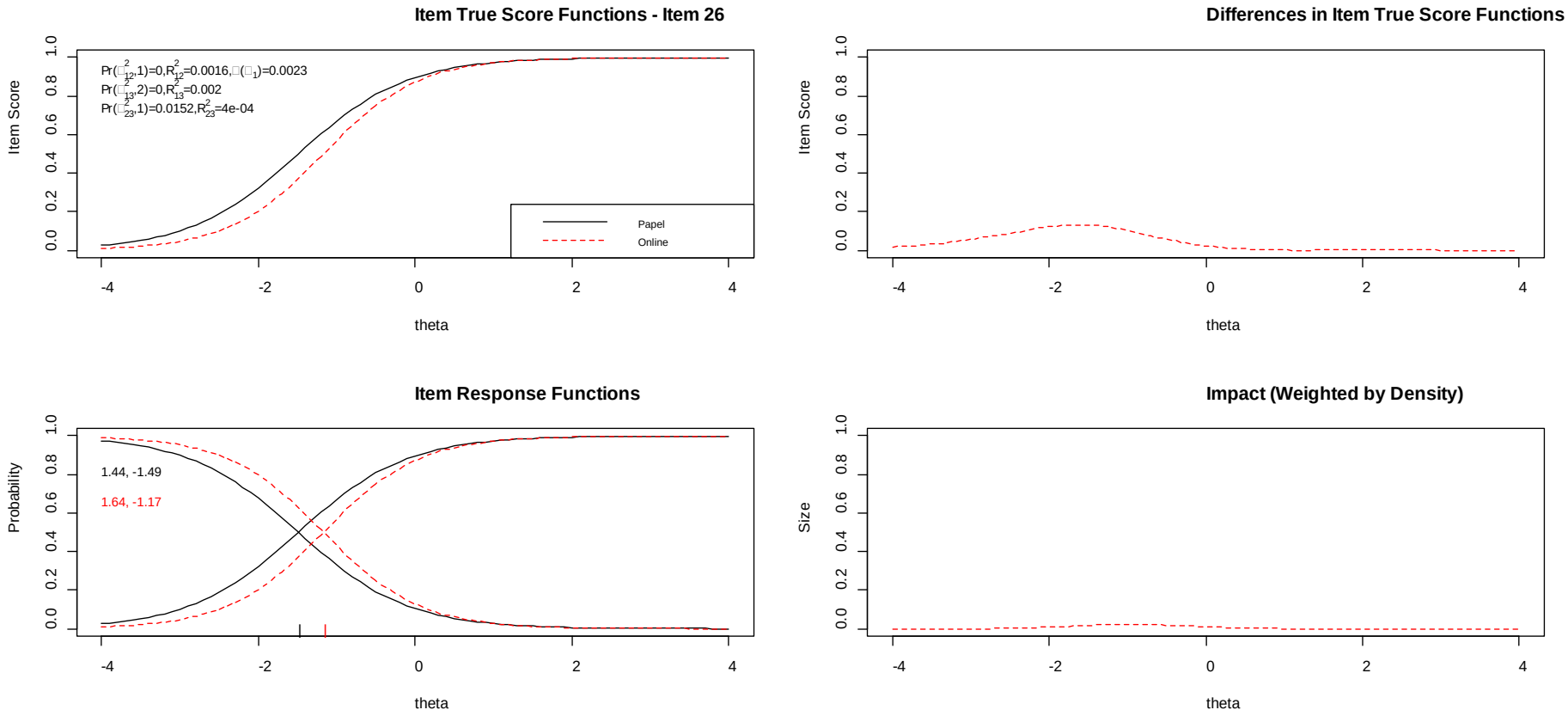


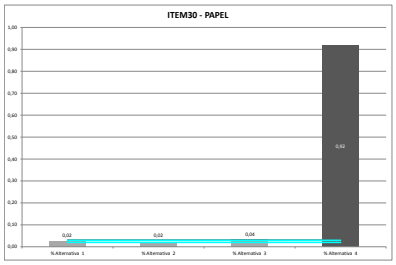
Figura 22. Características y Funcionamiento Diferencial de Versiones en el Ítem 26

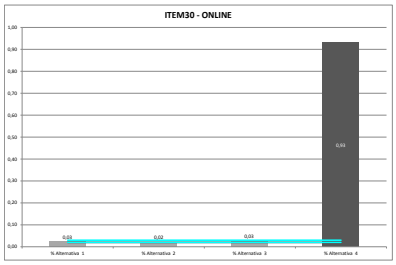
Tabla 19.
Características y Funcionamiento Diferencial de Versiones en el Ítem 30

Descripción desde la TCT

Ítem 30	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,92	0,07	0,27	0,46	0,42	4	24,5	15,8	0,91	0,29
Online	0,93	0,06	0,25	0,40	0,35	4	24,8	17,6	0,92	0,272

Porcentaje de elección de cada alternativa – ítem 30 en papel y online





Modelo TRI de 2 Parámetros

Ítem 30	Parámetro a	Parámetro b	p
Papel	1,498	-2,182	0,101
Online	2,116	-1,827	0,347

Técnicas detección DVF Ítem 30

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	DVF	DVF	DVF
0,000	0,000	0,058	Uniforme	0,0018	0,0021	0,0004	Débil					

*Diferencias significativas ($p \leq 0{,}0091$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

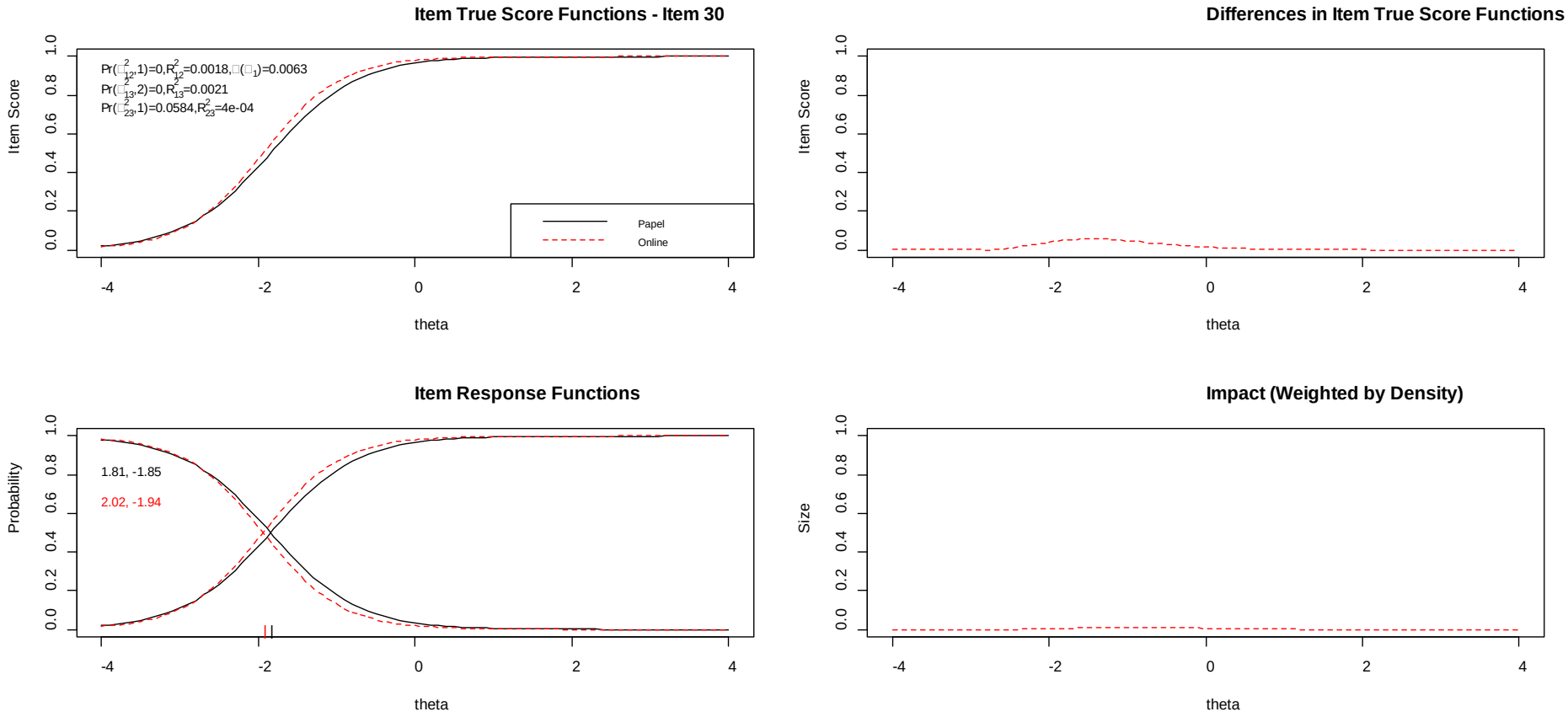


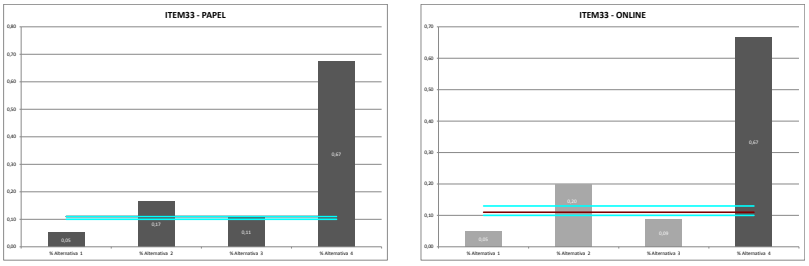
Figura 23. Características y Funcionamiento Diferencial de Versiones en el Ítem 30

Tabla 20.
Características y Funcionamiento Diferencial de Versiones en el Ítem 33

Descripción desde la TCT

Ítem 33	Facilidad	Varianza	Desviación típica	Biserial puntual	Biserial puntual corregida	Clave	Media p	Media q	Media	Desv. típ.
Papel	0,67	0,22	0,47	0,37	0,28	4	25,2	21,1	0,68	0,466
Online	0,67	0,22	0,47	0,38	0,29	4	25,6	21,9	0,65	0,479

Porcentaje de elección de cada alternativa – ítem 33 en papel y online



Modelo TRI de 2 Parámetros

Ítem 33	Parámetro a	Parámetro b	p
Papel	0,758	-1,150	0,686
Online	0,656	-0,814	0,614

Técnicas detección DVF Ítem 33

Regresión Logística								T.I.D.	Stand.	Raju	Lord	M-H
chi12	chi13	chi23	Tipo DVF	R^2_{12}	R^2_{13}	R^2_{23}	Efecto DVF	No DVF	No DVF	No DVF	No DVF	No DVF
0,019	0,003	0,014	No Uniforme	0,0003	0,0005	0,0003	Débil					

*Diferencias significativas ($p \leq 0,010$) nivel crítico corregido por Benjamini y Hochberg

Fuente: Elaboración propia

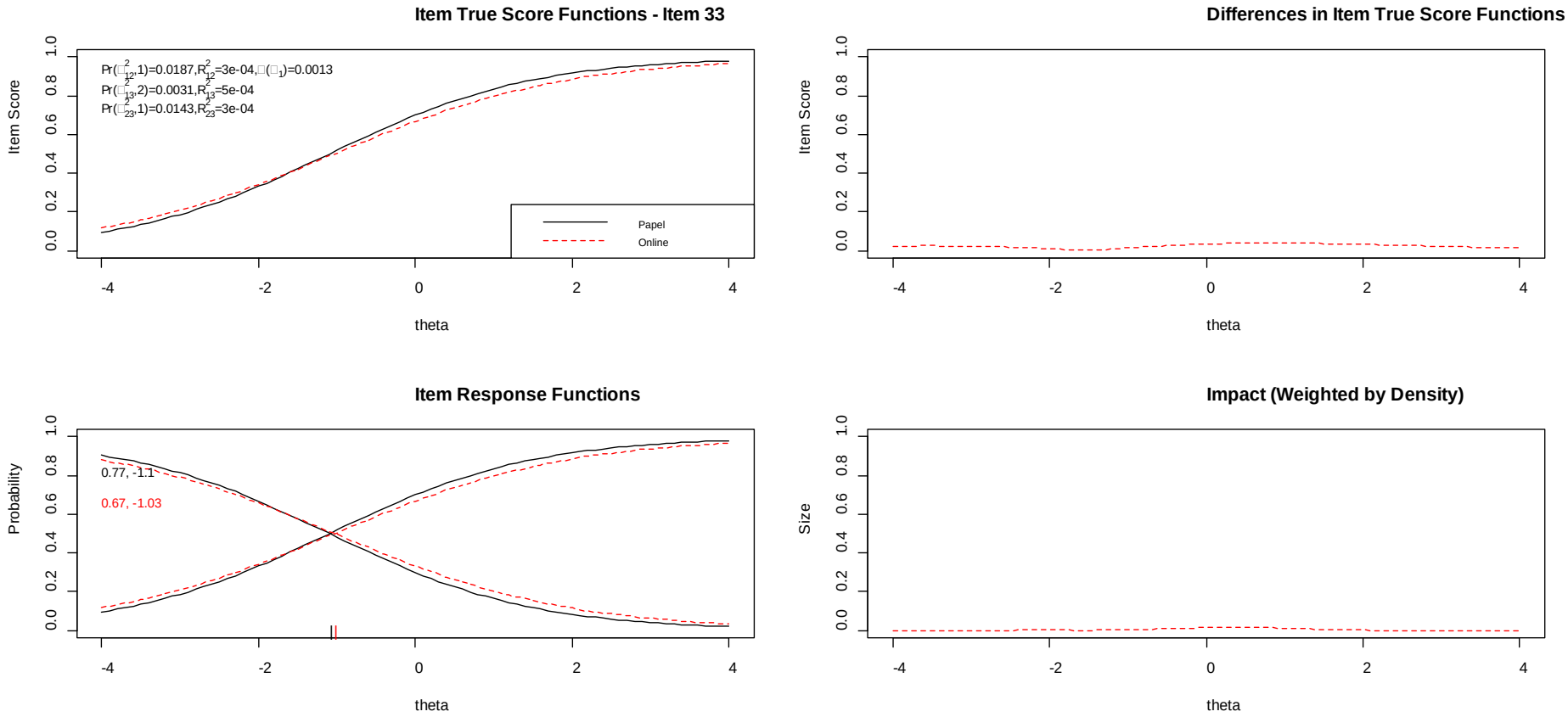


Figura 24. Características y Funcionamiento Diferencial de Versiones en el Ítem 33